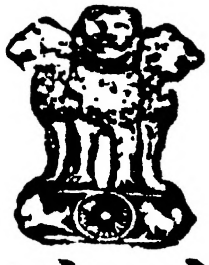


सत्यमेव जयते

I.A.R I.6.
GIP NLK—H 3 I.A.R.I.—10-5-55—15,000

Journal of Applied Physics

VOLUME 21

JANUARY–DECEMBER, 1950

Published by the

AMERICAN INSTITUTE OF PHYSICS

Incorporated

INFORMATION FOR CONTRIBUTORS

Suggestions for Contributors

Send manuscripts to Elmer Hutchisson, Case Institute of Technology, Cleveland 6, Ohio. Textual material (abstract, the body of the paper, tabular material, references, and figure captions) must be in English, typewritten (except mathematical expressions) double spaced with wide margins on a durable, white paper, preferably letter size. The original copy, not the carbon, should be submitted. For general style, spellings, abbreviations, and form, consult recent issues of this journal. Authors' institutions are requested to pay a publication charge of $8 per page which entitles them to 100 reprints without further charge.

An **abstract** must accompany each article. It should be adequate as an index and as a summary.

All but the simplest **mathematical expressions** should be carefully printed in by hand. Be sure to distinguish between capital and small letters. Indicate clearly superscripts and subscripts by inverted carets and carets, respectively. Avoid complicated superscripts and subscripts. Avoid repetition of a complicated expression by representing it by a symbol. Identify in the margin Greek letters and unusual symbols. Use fractional exponents instead of root signs. Use the solidus (/) for fractions wherever its use will save vertical space.

References should appear as footnotes, numbered consecutively, and arranged thus:

[1] A. B. Smith, Phy. Rev. 41, 852 (1932).
[2] H. Lamb, *Hydrodynamics* (Cambridge University Press, Teddington, England), sixth edition, pp. 573, 645.

Tables may be typewritten on sheets separate from the running text. Each table must have a caption which will make the data in the table intelligible without reference to the text. Avoid complicated column headings.

Wherever possible **figures** should be planned for the width of a column (3 in.). A convenient reduction ratio is about 2:1 or 3:1. Lettering should be large enough on the original that when reduced it shall not be less than about $\frac{1}{16}$ in. ($\sim$1.5 mm). Line drawings must be in *India ink* on white paper or tracing cloth. Coordinate paper is not desirable but if used must be blue lined and the coordinate lines which are to show must be inked over. Photographs should have maximum black and white contrast.

Identify each original drawing or photograph on the back with the author's name and the figure number. Figures should be numbered consecutively. Captions should be typed on a separate sheet of paper.

A limited number of **alterations** in proof are unavoidable, but the cost of making extensive alterations after the article has been set in type will be charged to the author.

Proof and all correspondence concerning papers in the process of publication should be addressed to the Publication Manager, American Institute of Physics, 57 East 55 Street, New York 22, New York.

INFORMATION FOR SUBSCRIBERS

The JOURNAL OF APPLIED PHYSICS is devoted to physics in its role as the science basic to other natural sciences and to the arts and industries. Previous to 1937 this publication was known as PHYSICS. It aims to be of service not only in physical laboratories, but also in the laboratories of industrial, chemical, geological, geophysical, meteorological, radio, and similar concerns. The JOURNAL OF APPLIED PHYSICS publishes editorials and reviews, as well as technical papers of applied physics.

Subscriptions and orders for back numbers should be addressed to the American Institute of Physics, 57 East 55 Street, New York 22, N. Y.

SUBSCRIPTION PRICE

	U. S. and Canada	Elsewhere
To members and associate members of the American Institute of Physics	$7.00	$8.00
To all others	9.00	10.00

BACK NUMBER PRICE

Complete set: Vol. 1, July, 1931—Vol. 19, 1949—$148.00
Yearly back number rate: $10.00.
Single copies: $1.00 each.

Advertising rates will be supplied on request.

American Institute of Physics

57 East 55 Street **New York 22, New York**

Journal
of
Applied Physics

Volume 21, Number 1 January, 1950

Thermionic Work Function of the (100) Face of a Tungsten Single Crystal

A. A. Brown,* L. J. Neelands,** and H. E. Farnsworth
Department of Physics, Brown University, Providence, Rhode Island
(Received July 5, 1949)

A slab with one surface cut and etched parallel to (100) crystallographic planes formed the cathode in a low power electron microscope. The electron image of the cathode was formed on a fluorescent screen with a small hole at the center. An electron collector placed behind the hole permitted a measure of the electron current passing through the hole as a function of the temperature of the cathode as measured with an optical pyrometer.

The most probable value of the work function, φ, obtained after 2400 hours of outgassing in a sealed-off, gettered tube is 4.59 ± 0.02 ev. This may be compared with 4.56 ev obtained previously by Nichols for the work function in the [100] direction of a tungsten single crystal wire.

INTRODUCTION

It is now quite well established experimentally that the work function of a metal single crystal is a function of the crystallographic plane which forms the surface boundary, although the amount of reliable experimental data is still very small. Some theoretical treatments[1] have been published but more experimental data are needed to check the validity of any satisfactory theory. Most of the previous measurements have been made photoelectrically or by means of contact potential differences, although Nichols[2] has obtained thermionic values for tungsten characteristic of certain crystal directions from measurements on a single crystal filament. Because these values may differ from those characteristic of the crystallographic planes of the same Miller indices, we have made measurements which are characteristic of the crystal planes. The (100) and (110) faces were chosen since they are the ones which appear most readily as a result of the heat treatment that tungsten filaments receive in use.

PREPARATION OF THE CRYSTAL

The crystal[3] was cut parallel to a specified set of crystal planes with an error of less than 0.05°, using a diamond wheel, and optical arrangement for alignment. The cutting marks left in the crystal surface by this method were removed by hand grinding. The resulting surface was substantially free of voids.

Attempts to remove the disturbed surface layer were made by (1) chemical etching with a number of different reagents and various conditions of temperature and concentration, (2) electrolytic etching, and (3) heat treatment. No single reagent or etching condition was found to expose smooth plane surfaces parallel to only one set of crystallographic planes. A (110) surface was prepared by first obtaining a good electrolytically polished surface,[4,5] and then heating the crystal, by electron bombardment of the back surface, at 2600°K. The resulting single crystal surface was quite smooth except for a few small pits corresponding to the natural voids of the crystal. A (100) surface was similarly prepared and heated to 2600°K for 100 hours. The resulting surface, although as smooth as the prepared (110) surface, was slightly more pitted. However, even in

* Now with Pratt and Whitney Aircraft, Inc., East Hartford, Connecticut.

** Now with General Electric Company, Electronics Park, Syracuse, New York.

*** Based on dissertations presented for the degree of Doctor of Philosophy in Brown University.

[1] R. Smoluchowski, Phys. Rev. **60**, 661 (1941). I. N. Stranski and R. Suhrmann, Ann. d. Physik 6, **1**, 153 (1947). C. Herring and M. H. Nichols, Rev. Mod. Phys. **21**, 185 (1949).

[2] M. H. Nichols, Phys. Rev. **57**, 297 (1940).

[3] The crystal was obtained from Professor A. Geotz of the California Institute of Technology. It had been obtained previously from the Lamp Department of the General Electric Company, Cleveland, Ohio. Analyses of similar ingots indicated that there was less than 0.05 percent of extraneous material present.

[4] W. C. Elmore, J. App. Phys. **10**, 724 (1939).

[5] J. M. Hughes and E. A. Coomes, Phys. Rev. **55**, 1138A (1939).

this case the pits covered only a very small fraction of the total area.

Two slabs with sides parallel to the (100) and (110) crystallographic planes respectively were cut from a large single crystal; these were cut to a size of 8×5×1 and 8×4×1 mm, respectively. One edge of each was beveled at an angle of 45° so that when the two were placed with the beveled edges together there resulted a slab 8 mm square and 1 mm thick with a division line between the (110) and (100) faces across the front. These were then ground and polished as described above.

The advantage of studying the two faces chosen is due to the fact that while the (110) surface appears most readily upon heat treatment,[6,7] the (100) face appears to have the lower work function[2,8] so that small pits showing other faces will not affect appreciably any observed current readings.

APPARATUS

A schematic diagram of the experimental tube is shown in Fig. 1. The crystals formed the cathode, Q, of a low power electron microscope. They were heated by electron bombardment from the filaments at P. The electrons emitted from the front surface were

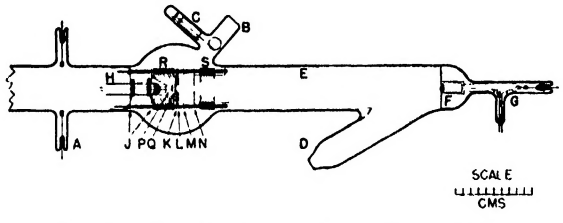


FIG. 1. Schematic diagram of experimental tube.

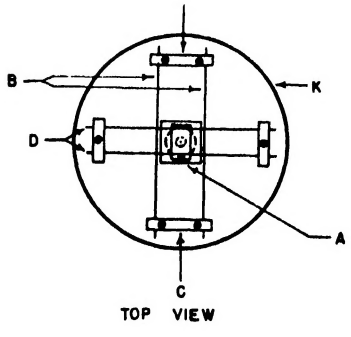
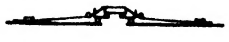


FIG. 2. Diagram of crystal mounting.



focused by means of the electrostatic lenses, L, M, and N, on the fluorescent screen and collector system, F. The temperatures of the crystals were measured by means of an optical pyrometer sighted through the side tube, B. In this tube a magnetically controlled shutter, C, prevented tungsten from depositing on the window at high crystal temperatures when readings were not being taken. The image could be deflected on the screen by a magnetic field and the emission current measured from any portion of either face by means of a Faraday collector placed behind a small aperture in the screen. R and S were guard rings used to prevent the passage of electrons from the bombarding filaments to the collector.

The crystals were attached to the diaphragm, K, by means of 0.5-mm tungsten wires passing through grooves in the backs of the crystals. A piece of 5-mil tungsten wire bent to form a 7-mm square was placed between the crystals and the diaphragm to decrease the heating of the diaphragm. During the later part of the work only one crystal was used. Consequently, the crystal mounting was changed to that shown in Fig. 2. In this mounting the heating of the diaphragm, K, was further decreased by interposing the small tungsten square A between the crystal and the diaphragm. This square was separated from both the crystal and the diaphragm by rectangles formed of 5-mil tungsten wire. The magnification produced by the electrostatic lenses was about ten.

The major part of the glass envelope is shown in Fig. 1. All glass and metal parts were thoroughly cleaned, in boiling dilute sodium hydroxide, in chromic acid, and rinsed in distilled water. The entire assembly was carried out using clean tools and sterile cotton gauze. No part was touched by the hands after being cleaned and a gauze mask was worn to prevent breathing on the parts.

Between the pumps and the experimental tube, and between the diffusion pumps and the fore pump, were traps cooled by dry ice and acetone. The tube, after being evacuated, was baked for two 24-hour periods. Between these bakings the crystals were heated 170 hours at a temperature of 1500°C. Just before the second baking, the trap between the pumps and the experimental tube was warmed to room temperature to remove any condensed vapors. After the second baking the ionization gauge indicated a pressure of 1.5×10^{-8} mm of Hg with the crystals cool and 6×10^{-8} mm of Hg with the crystals at a temperature of 1750°C. After an additional 380 hours of outgassing with the crystals at 1750°C, the first getter was flashed and the tube sealed off. Following this the pressure did not rise above 4×10^{-8} mm of Hg with the crystals at 1750°C and after 1500 hours of outgassing at this temperature it did not rise higher than 1.5×10^{-8} mm of Hg. The flashing of the second getter at this time did not cause any detectable change in the vacuum conditions.

[6] I. Langmuir, Phys. Rev. **22**, 357 (1923).
[7] C. J. Smithells, *Tungsten—Metallurgy, Properties, and Applications* (D. Van Nostrand Company, Inc., New York, 1936).
[8] S. T. Martin, Phys. Rev. **56**, 947 (1939).

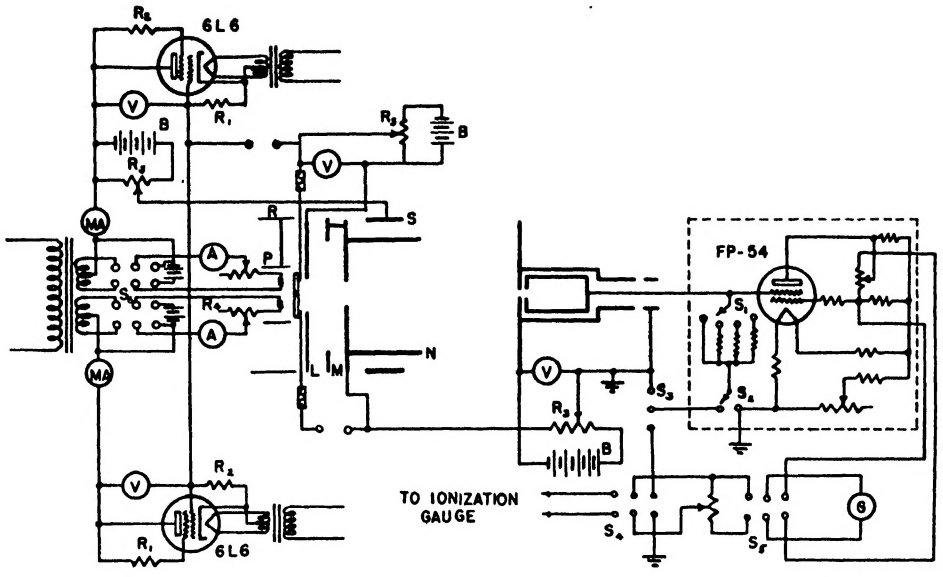


FIG. 3. Wiring diagram of the electrical circuit.

An optical pyrometer of the disappearing filament type was used to measure the temperature of the crystal surfaces. Experience indicated that even at temperatures as high as 2300°C, settings by different operators could be made as close as 2°C. The pyrometer was calibrated against a standard lamp obtained from the Bureau of Standards. The true temperature of the crystals was obtained from the brightness temperature by using the Wien equation and the value of the emissivity of tungsten obtained by Wahlin and Whitney[9] who found it constant and equal to 0.46 in the temperature range from 1200 to 2200°C at a wave-length of 6900A. Another correction was necessary due to reflection at the Pyrex window through which readings were taken. This correction was based on the work of Benford.[10]

A wiring diagram for the complete electrical circuit is shown in Fig. 3. The measuring system consisted of a d.c. amplifier and a galvanometer. This permitted the measurement of currents from 10^{-4} to 10^{-14} ampere.

PROCEDURE AND RESULTS

Measurements were started after 75 hours of heating and continued until the crystal had been degassed 2400 hours. All of the degassing except the first 240 hours was carried out at a temperature of 1750°C. The image was in focus when an accelerating potential of 2500 volts and a focusing potential of −200 volts with respect to the crystal were used. A retarding potential of 230 volts on the collector was sufficient to keep out the secondary electrons.

After the tube had been in operation for some time, evidence was obtained that some of the bombarding electrons were passing between the crystals and into the collector. This made measurements below 1300°C impossible. When a short circuit developed in the tube and made a reassembly necessary, the crystal mounting was redesigned and only one crystal with the (100) face was used. Consequently the measurements on the (110) face are of qualitative significance only, since they were taken before the most favorable conditions were obtained. The discussion which follows pertains only to the later work on the (100) face.

The measurements of the emission current and temperature were fitted to the formula

$$i = AT^2 \exp[-(\varphi/kT)]. \quad (1)$$

The value of φ was determined from the slope of the experimental curve obtained by plotting $\log i/T^2$ against $1/T$. It was not possible to calculate A since it was not ascertained that all of the electrons from the crystal passed through the various apertures.

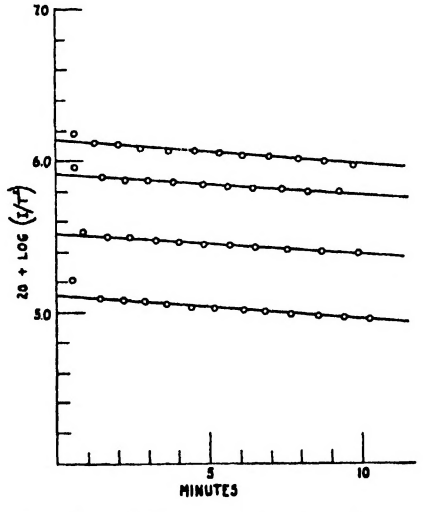


FIG. 4. Plot of $\log(i/T^2)$ against time elapsed after flashing the crystal at 2650°K for 15 seconds.

[9] Wahlin and Whitney, Phys. Rev. **50**, 735 (1936).
[10] F. Benford, J. Opt. Soc. Am. **29**, 162 (1939).

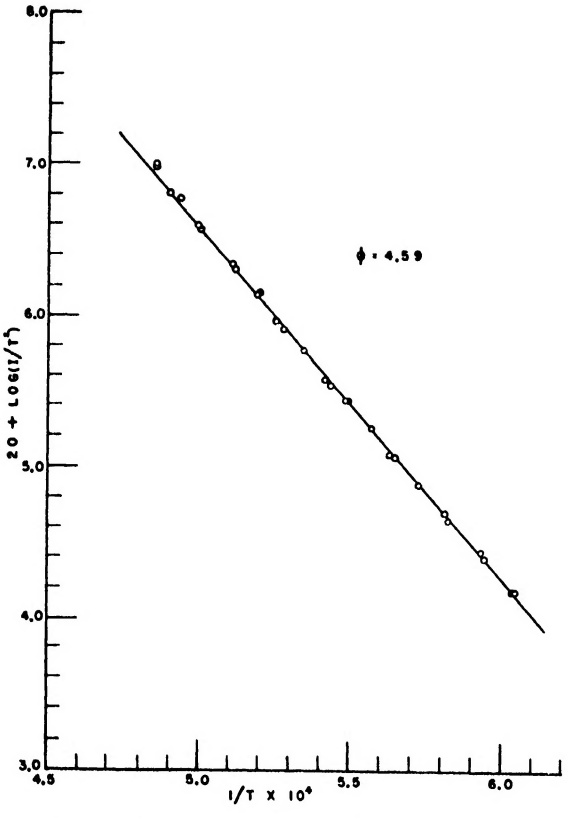


FIG. 5. A typical Richardson graph after 2000 hours of outgassing.

After about 900 hours of heating, the gas pressure as indicated by the ionization gauge did not continue to decrease with outgassing time and the procedure was adopted of flashing the crystal a short time at a temperature of 2650°K to obtain an uncontaminated surface. It was found that a flashing time of 5 sec. or longer resulted in a maximum emission current, and that a flashing time of 30 seconds did not increase the gas pressure over 4×10^{-8} mm of Hg. Consequently a flashing time of 15 seconds was used. Since the crystal surface became contaminated again in less time than was needed to complete a set of measurements, it was necessary to flash the crystal and then follow the emission current as a function of time and extrapolate back to zero time to obtain the values corresponding to the uncontaminated surface. The log i/T^2 was plotted which resulted in a set of straight parallel lines as shown in Fig. 4. The first one or two points are high in each case, since the measurements were made before the crystal had cooled to the desired temperature. This procedure was adopted after about 1000 hours of outgassing. The value of φ was then 4.77 ev. During the next 1000 hours of heating, the value decreased to 4.59 ev. It remained constant at this value until the work was ceased after a total outgassing time of 2400 hours. A typical plot, taken during the last 400 hours, is shown in Fig. 5.

For greatest accuracy, account should have been taken of the effect of the electric field at the crystal surface on the value of the work function. That is, every value of the current should have been corrected to the corresponding zero field value by using the relation

$$i_v = i_0 \exp(SV^{\frac{1}{2}}/T), \tag{2}$$

where i_v is the high field current; i_0 is the zero field current; V is the accelerating potential; and S is a constant which depends on the geometry of the apparatus. Equations (1) and (2) may be combined to give

$$i_v = AT^2 \exp[(\varphi - \Delta\varphi)/kT], \tag{3}$$

where $\Delta\varphi$, the correction due to the field, is given by

$$\Delta\varphi = KSV^{\frac{1}{2}}. \tag{4}$$

The configuration of the electric field in the microscope was such that the field was very low at the crystal surface and it was believed that no correction would be necessary. To confirm this, measurements of φ were made with the accelerating (and focusing) voltage varied by a factor of two. The values checked within the experimental error. It was found impractical to measure S by measuring i_v as a function of the accelerating potential at a constant crystal temperature, since this temperature could not be held constant for the length of time needed to complete a set of measurements.

CONCLUSIONS

The most probable value of φ for the (100) surface, obtained after 2400 hours of outgassing, was found to be 4.59 ± 0.02 ev. This is believed to be characteristic of the uncontaminated crystal surface. This checks well with the value of 4.56 ev obtained by Nichols[11] for the [100] direction. No definite value was obtained for the (110) face but the early results indicate that φ for this face is higher than φ for the (100) face. This is also true of Nichols' results for the [110] direction.

[11] The values of φ first published by Nichols are 0.03 ev to low due to a numerical error. Corrected values are given in Rev. Mod. Phys. **21**, 202 (1949).

Interference of Growing Spherical Precipitate Particles*

C. WERT AND C. ZENER
Institute for the Study of Metals, University of Chicago, Chicago, Illinois
(Received June 27, 1949)

The exact theory of the rate of growth of spherical precipitate particles has previously been developed for the initial stage where the individual precipitate particles do not interfere with one another. In the present paper this theory is carried one stage further to include their mutual interference. It is found that in the particular case of a very dilute solution this interference may be accounted for in a relatively simple manner. The applicability of the theory has been tested by new observations upon the precipitation of carbon from α-iron. Good agreement is found up to 95 percent completion of precipitation.

A. INTRODUCTION AND RESULTS

ONE of the authors (C. W.) has recently studied the rate of precipitation of carbon from freshly quenched α-iron.[1] From the observation that during the initial stage of precipitation the fraction of precipitation completed, W, obeys the equation

$$\tau dW/dt=(\tfrac{3}{2})(t/\tau)^{\frac{1}{2}}, \tag{1}$$

where τ is a constant, he concluded (a) that the growth nuclei were already present when the material was first quenched; (b) that the precipitate particles started as small spheres and remained spherical in shape throughout their growth. During the later stages of precipitation the growth of each precipitate particle is impeded by the general depletion of carbon in the surrounding matrix. An attempt was made to take account of this mutual interference of precipitate particles by the arbitrary introduction into the right side of Eq. (1) of an "interference factor" $(1-W)$. The observations were found to be in fairly good agreement with the resulting equation, namely,

$$\tau dW/dt=\tfrac{3}{2}(1-W)(t/\tau)^{\frac{1}{2}}, \tag{2}$$

which, when integrated, becomes

$$W=1-e^{-(t/\tau)^{\frac{3}{2}}}. \tag{3}$$

In spite of the fair success of this formula, the authors were unsatisfied with its heuristic derivation, and have since placed the theory of interference upon a firmer foundation. This new theory is developed in Section B. Experiments have been made to test the applicability of this theory to the particular system of carbon precipitating from solid solution in α-iron. These experiments are presented in Section C.

The results of the analysis of Section B may be presented in a simple form. Toward this end we rewrite Eq. (1) in the equivalent form

$$\tau dW/dt=\tfrac{3}{2}W^{\frac{1}{3}}. \tag{4}$$

It is then found that in the particular case of a very dilute solution the mutual interference of growing precipitate particles is properly taken into account by multiplying the right side of Eq. (4) by the interference factor $(1-W)$. The governing equation is, therefore, found to be

$$\tau dW/dt=(\tfrac{3}{2})(1-W)W^{\frac{1}{3}}. \tag{5}$$

While the original Eq. (2) can be directly integrated, the authors have been unable to integrate Eq. (5) analytically. However, W is a unique function of t/τ, and this function may be found by numerical integration. The results of such an integration are presented in Figs. 1 and 2. In these same figures a comparison is made with the solution of the original Eq. (2). It is seen that while both methods of taking interference into account lead to essentially the same results during the early stages of precipitation, they lead to markedly different results during the later stages. The very close agreement of the present theory with the observed precipitation is shown in Fig. 3. Tentative reasons for the deviations in the very last stages of precipitation are discussed in Section C.

B. ANALYSIS

The rate of growth of a spherical precipitate particle is completely determined by the gradient in concentra-

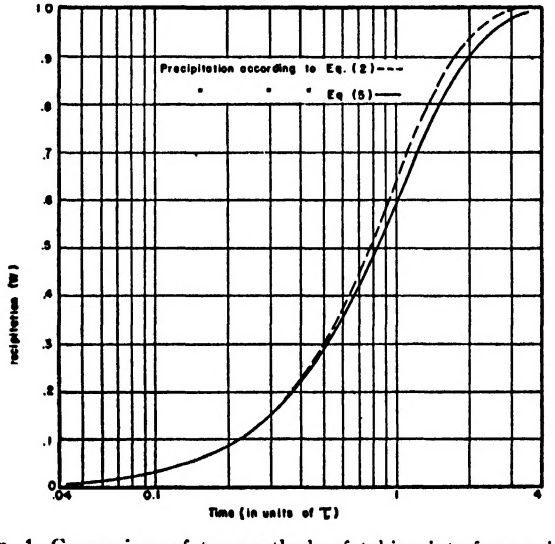


FIG. 1. Comparison of two methods of taking interference into account, standard coordinates being used.

* This research has been partially supported by the Office of Naval Research, U.S.N. (Contract No. N-6ori-20-IV, NR 019 302).
[1] C. Wert, J. App. Phys. 20, 943 (1949).

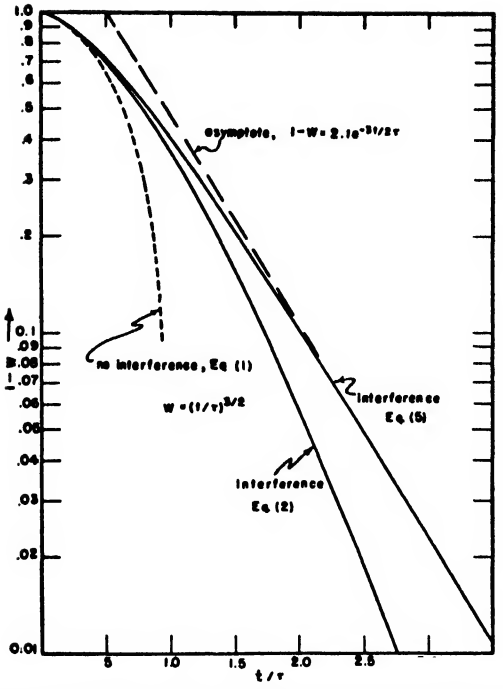


FIG. 2. Comparison of two methods of taking interference into account, coordinates being used which emphasize later stages of precipitation.

tion in the matrix adjacent to the particle. We shall let $n(r, t)$ denote the concentration of solute atoms in terms of number of atoms per unit volume. The value of n in the precipitate will be denoted by n_0, and the boundary value of n in the matrix at the precipitate-matrix interface will be denoted by n_1. The rate of growth is then given by

$$(n_0-n_1)(dV/dt)=4\pi R^2 D(\partial n/\partial r)_{r=R}, \qquad (6)$$

where V and R are the volume and radius of the particle, respectively, and D is the diffusion coefficient of the solute atoms. In the particular case of a very dilute solution the particle grows so slowly that the growth rate itself has a negligible effect[2] upon the gradient $(\partial n/\partial r)_{r=R}$. We may thus assign this gradient the steady-state value it would approach if the radius were to remain constant. The steady-state solution of the diffusion equation in the case of a fixed radius R is

$$n(r)=n_\infty-(n_\infty-n_1)(R/r),$$

and hence we shall equate the concentration gradient at the boundary to

$$(\partial n/\partial r)_{r=R}=(n_\infty-n_1)/R. \qquad (7)$$

This gradient is fortunately quite insensitive to the precise manner in which we specify the boundary conditions far away from the particle. The gradient is, e.g., not appreciably altered if we specify that the concentration be n_∞ at some radius R_∞ very large compared to R, or if we specify that $\partial n/\partial r$ vanish at some radius R_∞ large compared to R. In the later case n_∞ will be the concentration at R_∞, and will gradually decrease with time.

The assumption of a very dilute solution enables us to make one further simplification. In this case the precipitate particles will be relatively far apart, so that essentially each precipitate particle is surrounded by a concentration field which approaches the average concentration $n_\infty(t)$ as we go away from the particle in any direction. In the case of a dilute solution this average concentration is related to the fraction W of the precipitation which has already occurred by the equation

$$\{n_\infty(t)-n_1\}/\{n_\infty(0)-n_1\}=1-W. \qquad (8)$$

If we now let R_0 denote the final radius of the precipitate particles, corresponding to complete precipitation, and observe that

$$W=(R/R_0)^3, \qquad (9)$$

we find that substitution of Eqs. (7) and (8) into Eq. (6) leads to Eq. (5) of Section A, with the time of relaxation τ given by

$$\tau=[\{n_0-n_1\}/\{n_\infty(0)-n_1\}]R_0^2/2D. \qquad (10)$$

This time of relaxation may be expressed in alternative forms through the introduction of the density of precipitate particles, N, i.e., their number per unit volume, which satisfies the relation

$$(4\pi/3)R_0^3N(n_0-n_1)=n_\infty(0)-n_1. \qquad (11)$$

We have implicitly assumed that all the precipitate particles attain the same final radius. In accordance with the empirical observations of one of the authors,[1] it is legitimate, at least in the case of carbon in α-iron, to assume that all the precipitate particles start growing at the same time, namely, at the instant of quenching. It might appear, however, as if fluctuations in the density N would be reflected as fluctuations in the final radii. In order to demonstrate the invalidity of this objection in the case of very dilute solutions we shall eliminate the concentration from Eq. (10) by use of the relation (11), thereby obtaining

$$\tau=3/(8\pi NR_0D).$$

If we now introduce a measure λ of the mean distance apart of the precipitate particles by the definition

$$N=\lambda^{-3},$$

we obtain

$$\tau=(3\pi\lambda/8R_0)(\lambda^2/\pi^2D).$$

The second factor is seen to be the time of relaxation τ_λ for the smoothing out of fluctuations in concentration over a distance λ. Under our assumption of a very dilute solution the first factor is a very large number, and hence τ is very long compared to τ_λ. Thus although those regions in which N is unusually large will suffer an unusually rapid rate of precipitation, the excess

[2] C. Zener, J. App. Phys. **20**, 962 (1949).

rate of depletion of the solute atoms will, to a large extent, be compensated by the long range diffusion of solute atoms from surrounding regions. The average concentration $n_\infty(t)$ will thus remain almost constant in spite of possible fluctuations in N.

In order to have a formula whereby we may calculate N from an empirical value of τ, we eliminate R_0 from Eqs. (10) and (11), therby obtaining

$$N=(\tfrac{3}{4}\pi)2^{-\frac{1}{2}}[(n_0-n_1)/\{n_\infty(0)-n_1\}]^{\frac{1}{2}}(D\tau)^{-\frac{3}{2}}. \quad (12)$$

C. EXPERIMENT

The experiments made to test the validity of the foregoing analysis were carried out in a manner nearly the same as that described in an earlier paper by one of the authors.[1] The apparatus for preparing the specimens and making the internal friction measurements was that used previously. The only real difference between the past and present work is that the earlier work placed emphasis on the initial part of the precipitation process, whereas the present experiments emphasize the latter part of the process.

Precipitation measurements were carried out in the temperature range 85°C to 210°C. These measurements are compared with the theoretical curves in Fig. 3. In this figure the times τ' and τ (which differ only slightly) refer to the time constants of Eqs. (2) and (5) respectively. These constants τ' and τ were so chosen for each temperature that the experimental points coincided with the theoretical curves when the precipitation was about 60 percent complete. In order that simultaneous comparison could be made between experiment and the two theoretical curves, the scale factors for the two equations were so chosen as to make the curves coincide at 60 percent precipitation. It is clear from these plots that the theory deduced in the preceding section fits the experimental facts much better than the treatment expressed by Eq. (2). The experimental points show good agreement with Eq. (5) up to about 90 percent to 95 percent of the total precipitate.

The reason for the discrepancy during the last 5 percent of the precipitation is not known definitely. A possible reason is lack of homogeneity of the initial

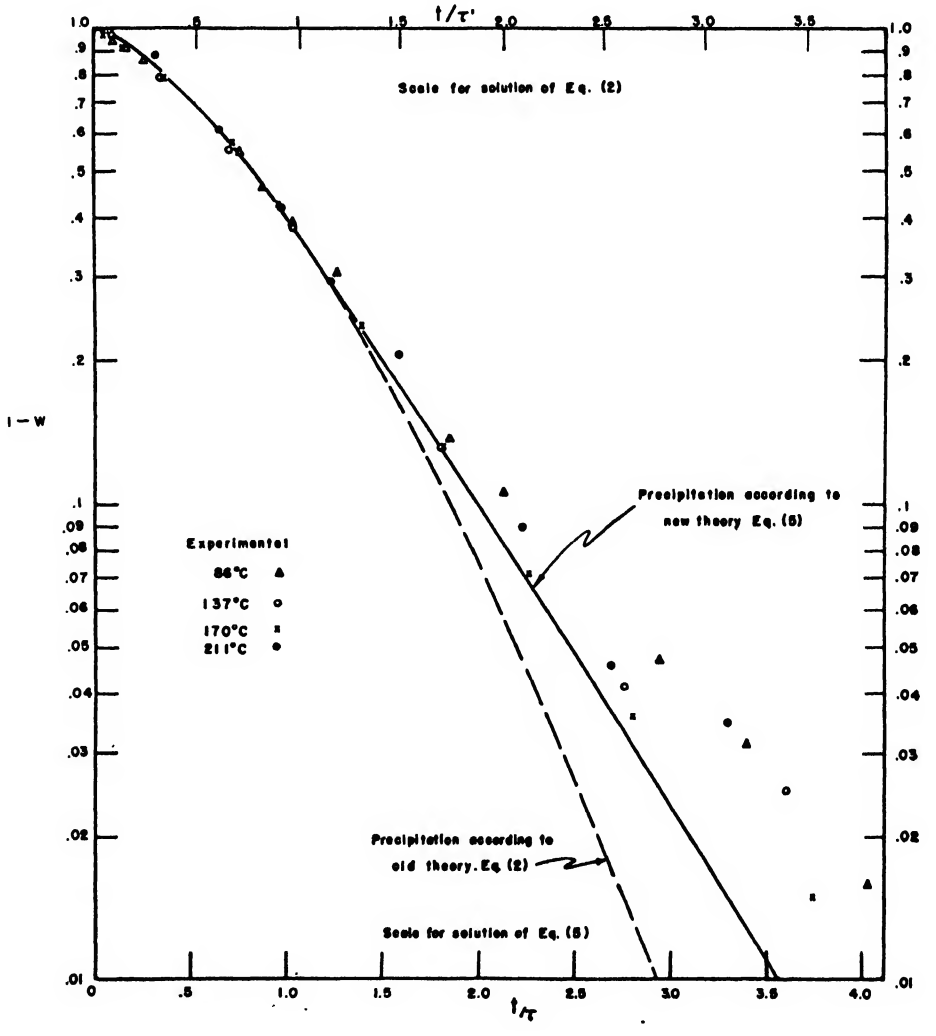


FIG. 3. Comparison of experiments with theory.

TABLE I.

Temp. °C	$n_\infty(0)$ (wt. percent)	τ(sec.)	N/cc
86	0.016	79000	39×10^{12}
137	0.013	3500	26×10^{13}
170	0.018	700	17×10^{13}
211	0.012	190	9×10^{13}

solid solution of carbon, which results in a range of values for τ. It was observed that specimens which were carburized without a long homogenization anneal gave very erratic and nonreproducible data. Further, the data on such specimens deviated from Eq. (4) as early as the 75 percent point of the total precipitation. When such a specimen was homogenized by an anneal of several hours at 720°, much more reproducible data were obtained. All the data shown in Fig. 3 were made on specimens which had been given such an anneal, and still longer anneals gave no improvement in the fit of the data to Eq. (4). However, it seems possible that slight inhomogeneities might still exist which would be large enough to account for at least a part of the discrepancies observed. A second possibility is the binding of dissolved carbon atoms to each other, to other dissolved impurity atoms or to lattice imperfections. That this binding may occur was observed experimentally earlier by the authors.[3]

Through the use of Eq. (12) and the experimentally determined values of τ, one can calculate the number of particles of precipitate per unit volume formed at a given temperature. Table I shows the result of such calculations. The values of N computed in this way are roughly the same as those computed earlier by the authors.[1,2]

[3] C. Wert and C. Zener, Phys. Rev. **76**, 1169 (1949).

The Response Characteristics of Linear Systems

MARCELLO CINI
Riv, Officine di Villar Perosa, Turin, Italy
(Received May 26, 1949)

For a linear system, steady-state response to sinusoidal stress cannot have a unique dynamic modulus and phase shift, but these quantities must be related to rate of creep; particularly they cannot be independent of external frequency.

A proof is given that such linear system is always equivalent to a viscoelastic system. Finally the discrepancy between the relations deduced and experimental results of various authors is pointed out and discussed.

ALTHOUGH hysteretic and elastic properties of rubber-like materials have been thoroughly investigated in the last years, no theory in good agreement with experimental results has been developed, at least as far as the author knows.

On one hand, very general and important results have been reached in the study of viscoelastic systems,[1–6] pointing out the various relationships which allow prediction of their behavior under given conditions of stress or strain; on the other hand many reports[7–14] have been published illustrating tests carried out on different stocks of natural and synthetic rubbers, and giving the derived values of dynamic modulus, loss factor, rate of creep and relaxation and similar quantities. In the present paper linear systems* will be examined from a very general point of view, and some fundamental properties of the response will be established, by means of which it is easy to prove that the most general linear system is equivalent to a viscoelastic model.**

Finally some experimental results of the above-mentioned papers are discussed using the derived relations and pointing out some reasons for the observed contradictions.

1. LINEAR SYSTEMS

We will assume as a fundamental property of a linear system that the steady-state response to a sinusoidal stress

$$P(t)=P_0e^{i\omega t}$$

[1] Fuoss and J. G. Kirkwood, J. Am. Chem. Soc. **63**, 385 (1941).
[2] Robert Simha, J. App. Phys. **13**, 201 (1942).
[3] T. Alfrey and P. Doty, J. App. Phys. **16**, 700 (1945).
[4] John G. Kirkwood, J. Chem. Phys. **14**, 180 (1946).
[5] B. Gross, J. App. Phys. **18**, 212 (1947).
[6] B. Gross, J. App. Phys. **19**, 257 (1948).
[7] Sack, Motz, and Work, J. App. Phys. **15**, 396 (1944).
[8] H. S. Sack and J. Motz, J. App. Phys. **18**, 450 (1947).
[9] Dillon, Prettyman, and Hall, J. App. Phys. **15**, 309 (1944).
[10] Kosten, Rubber Chem. Tech. **13**, 381 (1939).
[11] Stambaugh, Rubber Chem. Tech. **14**, 382 (1941).
[12] Oberto and Palandri, Rubber Age **63**, 725 (1948).
[13] Tobolsky, Prettyman, and Dillon, J. App. Phys. **15**, 380 (1944).
[14] Mooney, Wolstenholme, and Villars, J. App. Phys. **15**, 394 (1944).

* All the following considerations hold as well for electric, acoustic or dielectric systems.
** This identity is also implicitly proved in the papers of Alfrey and Doty (reference 3) and Gross (references 5 and 6.)

be still sinusoidal and given by

$$y(t)=[P_0/E(\omega)]e^{i(\omega t-\varphi(\omega))}.$$

If the sinusoidal stress is given in real form

$$P(t)=a\cos\omega t+b\sin\omega t$$

the response will be

$$y(t)=[a/E(\omega)]\cos(\omega t-\varphi(\omega))+[b/E(\omega)]\sin(\omega t-\varphi(\omega)).$$

As is well known $E(\omega)$ is the dynamic modulus, and $\varphi(\omega)$ the phase shift, both functions of the external frequency ω: $E(\omega)$ and $\varphi(\omega)$ are related to the mechanical impedance, or its reciprocal the mechanical admittance, by simple relations, that we need not report here.

This assumption follows directly from the principle of superposition, and is quite as general. Furthermore, many tests developed to investigate dynamic properties of rubberlike materials are based upon the analysis of steady-state response to sinusoidal stress,*** so that a rather accurate check of its shape is often possible.

Let us impose on such a system an external stress defined as follows:

$$\begin{aligned} P(t)&=0 \quad \text{for } t<0\\ P(t)&=P_0 \quad \text{for } 0<t<\theta\\ P(t)&=0 \quad \text{for } t>\theta. \end{aligned}$$

This function can be developed in Fourier's integral in a well-known way:

$$P(t)=\frac{P_0}{\pi}\int_0^\infty\left(\frac{\sin\omega\theta}{\omega}\cos\omega t+\frac{1-\cos\omega\theta}{\omega}\sin\omega t\right)d\omega.$$

Each infinitesimal harmonic of the stress, divided by the dynamic modulus $E(\omega)$ and shifted by the angle $\varphi(\omega)$, gives the corresponding infinitesimal harmonic of the response: integration from 0 to ∞ gives:†

$$y(t)=\frac{P_0}{\pi}\int_0^\infty\left[\left(\frac{\sin\omega\theta\cos\varphi}{E}-\frac{(1-\cos\omega\theta)\sin\varphi}{E}\right)\cos\omega t+\left(\frac{(1-\cos\omega\theta)\cos\varphi}{E}+\frac{\sin\omega\theta\sin\varphi}{E}\right)\sin\omega t\right]d\omega. \quad (1)$$

In order that Eq. (1) have a physical meaning two conditions must be satisfied:

(a) the response $y(t)$ must be independent of θ, i.e. from the time of application of external load, for $t<\theta$;

(b) the response $y(t)$ must be 0 for $t<0$, i.e. before the application of external load.

*** The presence of an inertial mass does not alter the shape of the response.

† In the following $E(\omega)$ and $\varphi(\omega)$ will be handled as E and φ, it being understood that they are functions of ω.

The first condition means that

$$\frac{\partial y}{\partial\theta}=0 \quad \text{for } t<\theta; \quad \frac{\partial y}{\partial\theta}\neq 0 \quad \text{for } t>\theta. \quad (2)$$

The derivative of Eq. (1) with respect to θ is:

$$\frac{\partial y}{\partial\theta}=\frac{P_0}{\pi}\int_0^\infty\left[\frac{\cos\varphi}{E}\cos\omega(\theta-t)-\frac{\sin\varphi}{E}\sin\omega(\theta-t)\right]d\omega.$$

Then, setting

$$\begin{aligned} \theta-t&=\tau \quad \text{when } t<\theta\\ t-\theta&=\tau \quad \text{when } t>\theta \end{aligned}$$

$$\int_0^\infty\frac{\cos\varphi}{E}\cos\omega\tau d\omega=\frac{\pi}{2}f_1(\tau)$$

$$\int_0^\infty\frac{\sin\varphi}{E}\sin\omega\tau d\omega=\frac{\pi}{2}f_2(\tau),$$

condition (a) is fulfilled if

$$f_1(\tau)=f_2(\tau)=f_1(-\tau)=-f_2(-\tau)=f(\tau),$$

i.e. if $f_1(\tau)$ is an even function, $f_2(\tau)$ an odd function, which are identical for positive values of the variable, being $f(\tau)$ their common value.

Inversion of Fourier transform gives

$$\frac{\cos\varphi}{E}=\int_0^\infty f(\tau)\cos\omega\tau d\tau \quad (3a)$$

$$\frac{\sin\varphi}{E}=\int_0^\infty f(\tau)\sin\omega\tau d\tau. \quad (3b)$$

It is easy to demonstrate that if Eqs. (3) hold, condition (b) is also satisfied. Indeed, introducing Eqs. (3) in Eq. (1), and inverting the order of integrations, we have

$$y(t)=\frac{P_0}{\pi}\int_0^\infty f(\tau)d\tau\int_0^\infty\left(\frac{\sin\omega(\theta-t+\tau)}{\omega}+\frac{\sin\omega(t-\tau)}{\omega}\right)d\omega. \quad (4)$$

Remembering that

$$\int_0^\infty\frac{\sin kx}{x}dx=\begin{cases}\frac{\pi}{2} & \text{for } k>0\\ -\frac{\pi}{2} & \text{for } k<0\end{cases}$$

we obtain

$$\int_0^\infty\left(\frac{\sin\omega(\theta-t+\tau)}{\omega}+\frac{\sin\omega(t-\tau)}{\omega}\right)d\omega=\begin{cases}0 & \text{for } t<0\\ \pi & \text{for } 0<\tau<t<\theta\\ 0 & \text{for } 0<t<\tau<\theta.\end{cases}$$

Then Eq. (4) can be written

$$y(t)=P_0\int_0^t f(\tau)d\tau \quad \text{for } 0<t<\theta$$
$$y(t)=0 \quad \text{for } t<0. \tag{5}$$

The derivative of (5) with respect to time gives

$$dy/dt=P_0f(t), \tag{6}$$

which shows that $f(t)$ is the rate of creep under unit stress.

Equations (3) and (5) do not depend on the magnitude of θ and therefore hold as general laws that must necessarily be fulfilled correlating dynamic modulus and phase shift (or real and imaginary part of mechanical impedance or admittance) with rate of creep. In identical way analogous relations between dynamic modulus, phase shift and rate of relaxation could be developed.

From Eqs. (3) it follows immediately that

$$\lim_{t\to\infty} f(t)=0 \tag{7}$$

$$\lim_{t\to 0} tg\varphi=0. \tag{8}$$

2. CORRELATION WITH VISCOELASTIC SYSTEMS

Identity of the given system with a convenient viscoelastic model, is also proved by means of Eqs. (3). Indeed it is not *a priori* quite plain that all systems whose response to stress obeys the superposition principle can be represented by a model constructed from a discrete set, or a continuous distribution, of elements which obey Hooke's law, and elements which obey Newton's viscosity law.

Remembering that creep under a constant stress P_0 of a viscoelastic system having a distribution function of retardation times $F(\tau)$ is

$$y(t)=P_0\beta\left(1-\int_0^\infty F(\tau)e^{-t/\tau}d\tau\right)$$

with

$$\int_0^\infty F(\tau)d\tau=1,$$

we have

$$\frac{dy}{dt}=P_0\beta\int_0^\infty \frac{F(\tau)}{\tau}e^{-t/\tau}d\tau.$$

As Eq. (6) gives the rate of creep for the linear system, the condition for identity of this one and the former one is

$$f(t)=\int_0^\infty \rho(s)e^{-ts}ds, \tag{9}$$

where

$$s=\frac{1}{\tau}-\beta\frac{F(1/s)}{s}=\rho(s).$$

This means that $f(t)$ is the Laplace transform of the auxiliary distribution function $\rho(s)$.

The determination of $\rho(s)$ is always possible because $f(t)$ must obey the conditions that assure evaluation of integrals in Eqs. (3) which are sufficient to allow the inversion of the Laplace transform.

As Eqs. (3) and (9) are identical to Eqs. (6) and (25) in Gross paper,[6] all further developments such as methods of evaluation of distribution functions, are omitted here, having already been published there.

3. DISCUSSION OF EXPERIMENTAL RESULTS

Many of the above mentioned reports agree in the statement that the elastic modulus $E(\omega)$ and the phase shift $\varphi(\omega)$ are with good approximation, independent of frequency ω in the range from a few cycles per sec. to about 1300 c.p.s. This property has also been assumed as a general law of internal friction replacing Newton's viscosity law, not only to take account of experimental results, but also because calculations can be carried out in a very simple way.[15, 16]

As Eqs. (3) and (8) show that this deduction is not compatible with the assumption of linearity, we must conclude as follows:

(a) Experimental results are not sufficiently extended and a different behavior of E and φ is to be expected in the ranges of extremely low and very high frequencies.‡

(b) Tested materials do not give a linear response; in this case E and φ would no longer have a meaning, but energy loss per cycle could still be measured.

(c) The presence of plastic deformations may alter the results in a considerable way independently of frequency.

These reasons may also coexist together, and, more or less, explain the observed results, but further tests should be undertaken to check them.

[15] Krall, Tech. Ital. I, 2 (1946).
[16] Cattaneo, Acta Pont. Acad. Sci. IX, 14 (1946).
‡ *Note added in proof.*—It has already been pointed out that E varies appreciably over a sufficiently large frequency range. See, for example, A. W. Nolle, J. App. Phys. **19**, 753 (1948).

The Use of Spraying Methods and of Volatile Suspending Media in the Preparation of Specimens for Electron Microscopy*

ROBERT C. BACKUS AND ROBLEY C. WILLIAMS
Department of Physics, University of Michigan, Ann Arbor, Michigan
(Received June 16, 1949)

Developments have been made in the use of spraying methods in the preparation of specimens for examination in the electron microscope. At the same time, two wholly volatile diluents, containing electrolytes and adjusted to normal *p*H, have been developed for use in forming the spray drops. The use of volatile electrolytes makes it unnecessary to wash the specimen after the droplet patterns have been formed on the specimen screens.

There are several qualitative and quantitative uses and advantages of the spray technique as described. The qualitative ones are these: (1) the brief drying time of the droplets helps to preserve the shapes of particles upon drying, and also allows studies to be made of rapidly reacting systems; (2) since the drop patterns are reproducibly representative samples of the suspension under investigation, qualitative assays of the particulate composition of the suspension can be made, and a comparison of the particulate composition of two closely similar suspensions can be made under ideal control conditions; (3) the droplet patterns are discretely bounded by blank areas of substrate film, and this fact makes practicable the detection of small and subtle differences between the fine structure of the specimen material and that of the substrate.

The quantitative uses of the spray technique are: (1) by use of reference particles of known numbers in a suspension it is readily possible to determine the volumes of droplets issuing from a spray-making device; (2) by use of reference particles it is also possible to make an assay of the number of other particles, such as a virus, per unit volume of mixed suspension; (3) if a suspension is highly purified, the particle weight of the material in suspension can be determined. Experiments involving the counting of about 10,000 particles show that the quantitative precision is as good as would be expected from the statistics of random sampling.

The construction and use of a spray gun are described, as well as the characteristics of two volatile diluents: ammonium acetate and ammonium carbonate.

I. INTRODUCTION

SPECIMENS for electron microscopy are commonly prepared by placing a drop of liquid containing some suspended material upon the film-covered specimen screen, and allowing the drop to dry either unassisted or hastened by the withdrawal of some of the liquid on a piece of filter paper. If the material is suspended in a non-volatile medium, such as one of the usual diluents containing salts, it is necessary to wash the microscope screen in distilled water subsequent to the initial drying of the suspension. It is generally agreed that this technique is unsatisfactory, owing to the chemical and physical changes that may occur during the drying period of several seconds to a few minutes, and to the inevitable non-uniform distribution of the suspended material as it dries on the screen. The exposure of biological materials of even moderate lability to an environment of high ionic strength, followed by immersion in distilled water, is a procedure likely to cause changes in the shapes and sizes of the specimen materials. Attempts have been made to alleviate some of these unfavorable conditions; notably by freeze-drying the specimen drop, and by treatment of the suspended material with a fixative such as formalin.

A method occasionally used since the earliest days of electron microscopy can be used to reduce many of the drawbacks of the usual procedure described above. This is a method by which the suspension is either allowed to settle on the specimen screen as tiny droplets, or is sprayed directly and forcibly upon the screen. As a result of this procedure, the screen is partially covered with discrete, roughly circular, droplet patterns from about 5 to 20 microns in diameter.

Spray methods have been described by Riedel and Ruska,[1] Haardick, Kausche and Ruska,[2] and by Cravath, Smith, Vinograd, and Wilson.[3] They appear not to have been put to any general use, however, probably because their advantages largely disappear when the specimen material is suspended in the non-volatile diluents commonly used. We have made developments in the spray technique with the primary purpose of applying it to the preparation of specimens of biological material, and have at the same time developed aqueous suspending media which are wholly volatile. The method now appears to be highly useful for both qualitative and quantitative electron microscopy. Its

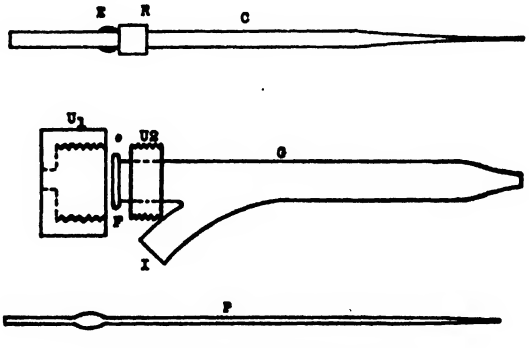


FIG. 1. Sketch of glass spray gun.

* This research has been supported in large part by a grant from the American Cancer Society upon recommendation of the Committee on Growth of the National Research Council, and by a research grant from the Division of Research Grants and Fellowships of the National Institute of Health, U. S. Public Health Service.

[1] G. Riedel and H. Ruska, Kolloid-Zeits. **96**, 86 (1941).
[2] Haardick, Kausche, and Ruska, Naturwiss. **27/39**, 226 (1944).
[3] Cravath, Smith, Vinograd, and Wilson, J. App. Phys. **17**, 309 (1946).

advantages and uses will be detailed below, following a description of a spray gun and a discussion of volatile suspending media.

II. CONSTRUCTION AND USE OF A SPRAY GUN

A spray gun should be capable of depositing upon the specimen screens a pattern of discrete droplets of approximate uniformity of size, reproducibly representative of the suspension, without the presence of contaminants originating in the gun, and devoid of any sprayed material in the regions between the droplet patterns. The simple glass gun to be described has been found adequate to meet these requirements, although it is realized that other gun designs might be found to be equally adequate. Figure 1 is a schematic drawing of the gun in which the parts are shown unassembled.

The outer envelope G is made of apparatus-weight, 7 mm glass, and is provided with an air inlet I and a means for securing the insert tube C. Tube G is provided with a threaded Lucite union U_1 and U_2, with U_2 of an internal diameter less than the diameter of the enlarged flange at F. U_1 threads over U_2, but also has an unthreaded base section with a hole just large enough to fit over tube C. The tube C is provided with a rubber sleeve R and a retaining bulge at E, and upon assembly an air-tight seal is made by compressing R between the flange F and the base of U_1. The purpose of E is to ensure that the tube C will not be violently forced out of the end of G when the compressed air is admitted. Tube C is of uniform diameter except for the bulge at E and the finely drawn portion at the tip.

It has been found convenient to make a number of small pipettes P whenever a series of samples is to be sprayed. The use of these pipettes minimizes the chances of sample-to-sample contamination by restricting the area of contact between fluid and glass to the small region at the tip of C. The pipettes are inserted into C sufficiently far to cause their ends to wedge lightly into the capillary constriction in C. Each pipette is discarded after having been used with one sample of a series.

Extended trial has shown that the dimensions of the orifices are fairly critical. The I.D. of tube C should be about 0.1 mm, and the O.D. about 0.3 mm. The I.D. of the tip of tube G should be about 0.75 mm. A spraying pressure of 20 to 30 p.s.i. is satisfactory, with the gun some 15 to 20 inches from the filmed microscope screens. The screens can be mounted to a supporting surface with Scotch tape, or can be held in position magnetically. Very clean air must be used for spraying, and it appears that tank-compressed air, reduced with a suitable valve, is satisfactorily clean. The gun should be carefully aligned to cause the central region of the spray to intersect the screens; the alignment and other spray characteristics can conveniently be observed with the aid of a beam of light in a darkened room. The gun as described will deposit a convenient number of droplets from about 0.2 ml of sprayed material. A smaller orifice can be used at the tip of tube C (with higher air pressure and diminished distance from gun to screens), when the use of a smaller volume of suspension is necessary.

One of the commercially available all-glass nebulizers has been experimented with, but it has not been found to be as satisfactory for the production of representative drop patterns as is the high air-velocity gun just described. The nebulizer is not as conveniently kept free of contamination, and it does not produce drop patterns that are both discretely bounded and reproducibly representative of the specimen. We believe that the reason for the difference in the performance of the two kinds of spraying devices is that, in the case of the low air-velocity nebulizer, some material from droplets which have completely dried in transit is left in varying amounts and places upon the microscope screens.

III. VOLATILE SUSPENDING MEDIA

The preparation of materials in the form of spray-drop patterns loses most of its unique value when the suspending medium is non-volatile, since the dried salts must be rinsed off the specimen screens with consequent disturbance of the drop patterns. For many substances a satisfactory volatile medium is distilled water, but for others, particularly of a biological nature, a medium containing electrolytes is demanded. The requirements to be met by volatile electrolytes when used in electron microscopy can be stated only after consideration of the type of observation to be made. A minimum requirement would be that the chemical and physical properties of the substance, when suspended in one of the conventional diluents, should be maintained when the substance is suspended in the volatile diluent.

It is clearly necessary that the preparation of a volatile electrolyte be done in a manner calculated to avoid any chance of contamination with non-volatile material. If a non-volatile component is present as much as 1 part in 10^7 in an average spray droplet of 500 microns^3 volume, there will be 5×10^7 A^3 of residual material in the drop pattern. This amount is clearly visible in the electron microscope if it dries on the screen in a form other than a rather uniform film. Such a high degree of freedom from non-volatile contamination requires that only gaseous or freshly redistilled liquid components be used in the chemical preparation.

We have found a 2 percent solution of freshly made ammonium acetate at pH 7.0 to be a satisfactory suspending medium for use with several biological materials (for example: fibrinogen, serum globulins, plant viruses, and homogenized tissue suspensions). Ammonium carbonate at a similar ionic strength and pH has also been found to be useful. It is advantageous to consider spraying the same material suspended in both of these volatile electrolytes in turn, on account of the changes in pH that occur during the brief drying time

of the droplets. The ammonium acetate system will approach a *p*H of about 4.0 as it dries, while ammonium carbonate approaches a *p*H of 10. If differences of form observable in the electron microscope are believed likely to take place in a drying time of about 0.1 sec., their magnitude can be checked by spraying with each volatile system separately, or with the two systems used together in varying concentrations.

IV. USES AND ADVANTAGES OF THE SPRAY METHOD

The consequences of the use of droplets sprayed in a volatile suspending medium may be generally described thus: (1) the time taken for the specimen material to change from its free liquid environment to complete dryness is very short, (2) the droplet patterns are individually and reproducibly representative of the composition of the sprayed material, and (3) the droplet patterns are discretely bounded by areas of blank specimen film. The uses and advantages accruing from these consequences will now be described in detail.

A. Qualitative

1. *Short drying time*—The total time elapsed between the emergence of an aqueous drop of about 500 microns[3] volume from a high air-velocity spray gun and the state of complete dryness on the specimen screen is estimated to be somewhat less than 0.1 sec. The estimate is based upon the known rate of evaporation of water droplets in relatively dry air, and disregards any possible effect of the suspended material in slowing or accelerating the drying process. This drying time is to be compared with a time of several seconds to a minute or so taken by an ordinary, gross drop to dry when placed directly upon the filmed specimen screen.

The advantage of the short drying time is that the suspended material, such as a virus, for example, is exposed only very briefly to an unfavorable electrolytic and osmotic environment and has little time to be morphologically changed. There is also little time for the occurrence of gross aggregation or other particle interactions which must take place during the concentration of the solute materials during the long drying time of a large drop. A further advantage is that very small particles do not imbed themselves in the plastic film as much as they do when the film is softened and swollen by prolonged contact with a large drop of water. An important use of the fact of quick drying is in observations of interparticle reactions, where it is desirable to stop the reaction quickly and at a known instant. In this type of experiment the particulate components are mixed prior to, or during, their introduction into the gun, and the reaction is manifestly stopped within a fraction of a second after the spray drops leave the gun.

2. *Reproducibly representative fields*—It is difficult to overestimate the usefulness of a technique whereby well-defined areas of the microscope screen contain representative and complete samples of all the non-volatile material in the original suspension. This is particularly true when the suspending diluent is in itself free of non-volatile components. The common experience of microscopists is that searching for and photographing "typical" fields is time-consuming, frustrating, and is open to the dangers of subjective judgment. This difficulty has made comparisons between two samples of material, such as normal and infected biological material, relatively uncertain, with the result that the electron microscope has failed to be used to its fullest as an analytical tool. The use of drop patterns eliminates a fundamental difficulty in electron microscopy; namely, the small totally available field of search of about 0.25 mm^2. This restricted field leads to uncertainty as to whether or not the portion of the specimen of great significance lies outside the field, or is obscured by the wires of the specimen grid.

A few qualitative uses of observations on small, representative drop patterns are immediately apparent. The first use lies in making reliable assays of the particulate composition of a given suspension by photographing a few drop patterns. Assays can be made by even an inexperienced person, since the choice of what is typical is independent of his judgment. If the suspension is believed to be a fairly homogeneous one, such as a suspension of a purified plant virus, the microscope can also be used to assay sensitively the amount of particulate impurities present.[4]

The spray method allows a precise comparison to be made of the electron microscopic appearance of two different, but similarly prepared, suspensions. The comparison can be conveniently made in this way: Suspension *A* is sprayed upon a few filmed specimen screens. Suspension *B* is adjusted to have about the same concentration of suspended material as *A*, but to it is added a trace amount of an indicator substance of recognizable shape, such as tobacco mosaic virus. Suspension *B* is then sprayed upon the screens already containing the dried droplets of suspension *A*. The screens are shadowed and examined for regions where discrete, closely adjacent drop patterns of both *A* and *B* can be photographed. This technique seems to embrace all the requirements of controlled electron microscopy, since the methods of mounting the specimen materials are closely identical, even to the details of drying time and surface character of the substrate film, as well as film thickness, shadowing conditions, and circumstances of exposure in the microscope.

3. *Discretely bounded drop patterns*—A constant source of uncertainty in the examination of shadowed electron micrographs of inhomogeneous particulate material is the difficulty of distinguishing between the fine-structure of the specimen material itself and the fine-structure of the substrate film. A similar difficulty

[4] R. C. Williams and R. C. Backus, J. Am. Chem. Soc. (in press).

is present even when preshadowed replicas[5] are used, except that the source of uncertainty is not in the structure of the collodion surface but instead resides in the uncertain degree of cleanliness and discontinuities of the glass surface. When drop patterns are examined, the uncertainty largely disappears, since the patterns occupy discrete areas surrounded by areas of film whose undisturbed structure can be gauged. This separation of regions provides one with a sensitive internal control on the character and cleanliness of the surface of the substrate, and hence enhances the possibility of

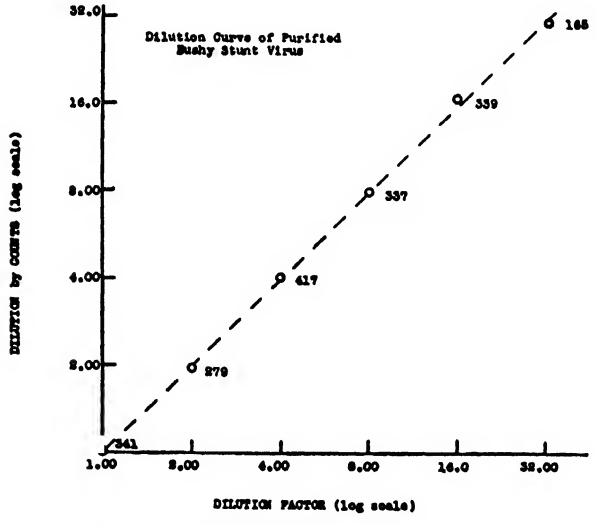


FIG. 2. Determination of the dilution of a suspension of purified bushy stunt virus by spraying and counting. Abscissae are the known dilution factors, obtained by pipetting; ordinates are the dilution factors calculated from the virus counts. The point at the origin has been arbitrarily called a relative dilution of unity. The diameter of the circles represents a 3 percent variation, while the numbers accompanying each circle are the numbers of reference particles counted in the establishment of each point.

making unequivocal observations of the presence or absence of very small particles.

B. Quantitative

The attainment of representative sample fields makes electron microscopy increasingly applicable to certain kinds of quantitative determinations. A technique to provide such fields by centrifugal deposition has been suggested by Crane.[6] A quantitative application of this technique has been made by Sharp[7] who obtained representative fields of swine influenza virus by sedimenting the particles on to a collodion surface. By counting the number of particles per unit area of the film he obtained a satisfactory correlation between counts and known dilution factors, although his results were rendered somewhat uncertain by the necessity of washing the dried buffer salts from the sedimented preparations.

[5] R. C. Williams and R. W. G. Wyckoff, J. App. Phys. **17**, 23 (1946).
[6] H. R. Crane, Rev. Sci. Inst. **15**, 253 (1944).
[7] D. G. Sharp, Proc. Soc. Exp. Biol. Med. **70**, 54 (1949).

1. *Determination of droplet volumes*—If reference particles[8] are present in a known concentration in the suspension being sprayed, a given droplet pattern will contain a number of these particles proportional to the volume of the droplet as it left the spray gun. It has been found that Dow latex particles[9] are particularly useful as reference particles, since they are obtainable in quite pure aqueous suspension, and remain intact and in uniform suspension in ammonium acetate and ammonium carbonate. The particles do not seem to interact with the biological materials we have handled, at least during the brief periods of contact necessary for their use, and their concentration in numbers per unit volume of suspension is readily determinable.[4] It is evident that one can use the latex reference particles directly in a determination of volumes of droplets as they issue from any spray-making device.

2. *Absolute virus assays*—When latex particles are mixed with some other particulate material, such as a virus, simple counts made on the droplet patterns will determine the ratio of the number of latex to the number of virus particles. If the concentration of latex is known, this ratio provides an absolute assay of the number of virus particles per unit volume of the mixed suspension. It is to be noted that this method can be used even though the suspension is not highly purified; as long as one species of particle can be recognized, it can be counted and assayed.

3. *Determination of particle weights*—The technique just described can be used to determine the particle weight of a homogeneous substance, such as a purified virus. The particle counts yield the number of particles per unit volume, and if the material is in highly purified form, this information can be combined with the value of the dry weight of material in an aliquot volume of suspension to allow the weight of an average particle to be computed. This weight in grams, multiplied by Avogadro's number, is the particle weight in molecular weight units. The advantage of this method of determining particle or macromolecular weights lies primarily in its directness, and also in the low concentration of material which can be sprayed and counted.* A satisfactory concentration is 0.0001 percent, and the total weight of material to be diluted and sprayed need be no more than 0.1 microgram.

4. *Assessment of quantitative precision*—Determinations have been made of the quantitative reliability and precision of the method of particle counts by the use of

[8] E. Chamot and C. W. Mason, *Handbook of Chemical Microscopy* (John Wiley and Sons, Inc., New York, 1938), Vol. I, p. 441.
[9] R. C. Backus and R. C. Williams, J. App. Phys. **20**, 224 (1949).
* When very dilute material consisting of small sized particles is sprayed, it is difficult to recognize the confines of the drop patterns in the electron microscope. The admixture of a small amount of highly purified tobacco mosaic virus to the suspension causes the drops to be nicely bounded by easily observed rod-like forms.

reference particles in the spray drops. Two sources of uncertainty might be anticipated: the mixing of the two kinds of particles may not be perfect, and there may be dislodgments of either species of particle from the microscope screens prior to shadowing.

The sources of uncertainty have been checked by measuring the statistical fluctuations of the counts, in a trial where particles of latex and of the bushy stunt virus were mixed in approximately equal numbers. It is found from counts of about 10,000 particles that the standard deviation of the mean ratio of particle numbers lies between 2 and 4 percent when 1000 pairs of particles are counted in 20 or more droplet patterns. Inasmuch as the statistically anticipated standard deviation of the mean ratio should be close to 3 percent, it is believed that the mixing of the particles is nearly perfect, and that they do not become randomly dislodged.

The adequacy of the entire method of quantitative assay of virus particles by electron microscopy has been checked by counts made on mixtures of reference and virus particles of greatly varying relative concentrations. Figure 2 shows the results of an experiment in which the concentration of latex particles was held constant, while the concentration of purified bushy stunt virus was varied through a dilution factor of 32. The evident agreement between the known dilution factors and the dilutions as calculated from the counts of virus and latex particles is very satisfactory, and implies that a quantitative assay of a purified, dilute virus suspension can be made as precise as the statistics of random sampling would indicate.

On the Diffraction of Radar Waves by a Semi-Infinite Conducting Screen*

C. W. Horton and R. B. Watson
Defense Research Laboratory, The University of Texas, Austin, Texas
(Received May 27, 1949)

The diffraction patterns of radar waves about a semi-infinite copper screen have been measured. The measurements were made by revolving a bisectoral horn in a circle of radius 15″ or 30″ about the diffracting edge. Measurements were made at three positions along the edge, one of which coincided with the corner of the screen so that the pattern obtained was actually that of a quarter-infinite plane. Measurements were made for two orientations of the screen corresponding to angles of incidence of 0° and −22½°. Theoretical patterns have been calculated for the entire 360° of rotation and these agree well with the experimental patterns. A comparison is made between the theory, measurements, and the equivalent optical measurements.

I. INTRODUCTION

A RIGOROUS solution of the problem of the diffraction of an electromagnetic wave by a semi-infinite perfectly-conducting screen has been obtained by Sommerfeld.[1,2] Optical measurements have shown deviations from the theory[3] which have usually been attributed to the finite thickness and conductivity of the screen and the irregularities of the edge of the screen.[4] All three of these difficulties are greatly reduced when the experiments are repeated at radar frequencies. A further advantage of the use of radar frequencies is that the smallness of the detector enables one to make significant measurements throughout nearly a full circle. Thus the theoretical results may be checked not only in the shadow region but also in the illuminated regions and especially in the region of geometric reflection. The disadvantages of the change in scale are (1) that a screen of manageable size is not large enough to eliminate diffraction around the other three edges and (2) that the effective aperture of the receiver is of appreciable area, though of the same order as the wave-length squared.

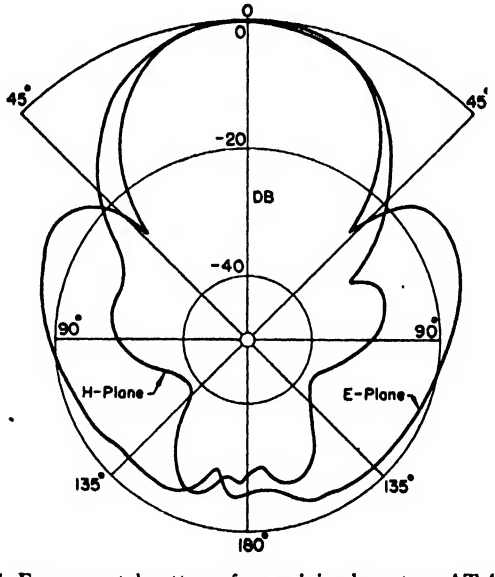


FIG. 1. Experimental patterns for receiving horn type AT-48/UP.

II. EXPERIMENTAL ARRANGEMENT

The experimental arrangement used in this work has been described previously.[5] A modification of this method has been made in which the receiving horn was mounted on a long arm so that it moved on a circle whose center coincided with the edge of the diffracting screen and which lay in a plane normal to the screen. The screen consisted of a single sheet of copper 30″×60″ and $\frac{1}{16}$″ thick which was mounted 30″ above the experimental bench with the long dimension horizontal. One of the 30″ vertical edges was the diffracting edge. The screen was supported by two U-shaped dural channels 2″×1$\frac{3}{8}$″ which were placed symmetrically about the vertical centerline. The channel nearer to the diffracting edge was 18″ from it. Small dural angles $\frac{3}{4}$″×$\frac{3}{8}$″ were required in order to stiffen the copper sheet and prevent vibrations since the sheet was mounted in the open air; however a space of 4$\frac{1}{2}$″ was left completely clear next to the diffracting edge. Some of these stiffeners were placed along the upper and lower edges of the screen.

The receiving horn was a bisectoral horn of the Type AT-48/UP manufactured for the armed services. The aperture of this horn is 1.9″×2.4″ or 1.47λ×1.82λ for a free space wave-length of 3.2 cm. Figure 1 shows the patterns of this horn in the *E*- and *H*-planes.

In order to observe the influence of the upper and lower reinforced edges, the free edge on the far end of the sheet, and the various supporting members, measurements were made with the receiving horn at two radii, 15″ (*ca.* 12 wave-lengths) and 30″ (*ca.* 24 wave-lengths), and at three positions along the diffracting edge. One position was at the midpoint, or 15″ from the upper or lower edges, a second position was 9″ from the upper edge, while the third position was at the upper edge of the screen. This last position really gives

* The work described in this paper was done at the Defense Research Laboratory under the sponsorship of the Bureau of Ordnance, Navy Department, Contract NOrd-9195.

[1] G. Wolfsohn, *Handbuch der Physik* (Julius Springer, Berlin, 1928), Vol. 20, pp. 266–277.

[2] B. B. Baker and E. T. Copson, *The Mathematical Theory of Huygen's Principle* (Oxford University Press, London, 1939), pp. 124–149.

[3] See reference 1, Figs. 6a and 6b, p. 275.

[4] See reference 1, p. 275.

[5] R. B. Watson and C. W. Horton, J. App. Phys. 19, 661–70 (1948).

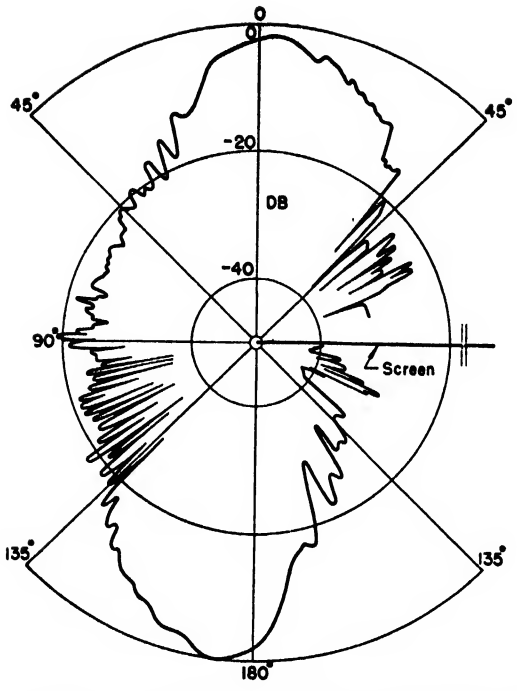


FIG. 2. Experimental diffraction pattern for semi-infinite screen for normal incidence, with incident wave from 0° direction. Receiving horn rotated in 30″ circle about center of diffracting edge. Electric vector parallel to diffracting edge.

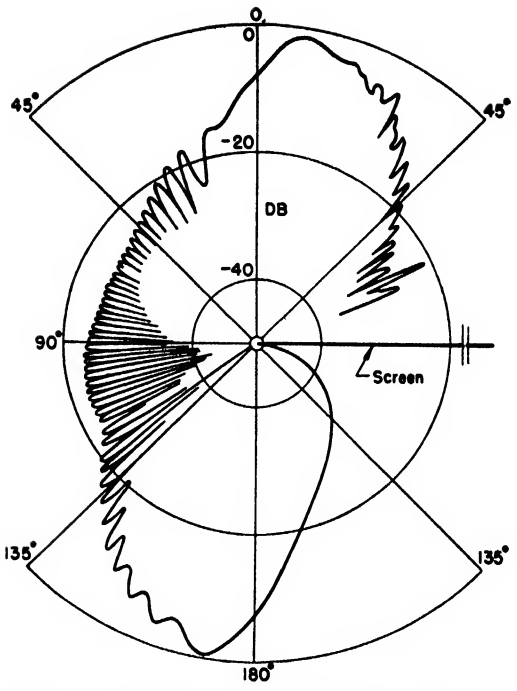


FIG. 3. Theoretical diffraction pattern for the conditions of Fig. 2.

patterns for a quarter-infinite screen. The agreement between the patterns measured about the center of the diffracting edge and the position 9″ from the upper edge is good. However, the presence of an interference pattern in the shadow region behind the screen for both of these positions shows that the screen was only a moderately good approximation to a semi-infinite screen.

Measurements were made for angles of incidence of 0° and of $-22\frac{1}{2}°$.

III. NORMAL INCIDENCE

Figure 2 is a polar plot of the received signal in db *versus* the angle about the edge of the semi-infinite screen for the case of an incident wave polarized with the electric vector parallel to the diffracting edge. The horn was always turned so that it faced the edge of the screen. Thus the diffracted wave was detected at maximum sensitivity and the directionality of the horn was used to discriminate against the incident and reflected waves. Figure 3 is a plot of the theoretical pattern for the same conditions. In computing the theoretical pattern the wave field was resolved into plane and cylindrical waves in accordance with the description of Baker and Copson.[6] The pattern of the horn was applied separately to each wave component in accordance with its orientation with respect to the horn. In the regions that are very near the line of propagation, the approximate expressions for the diffracted field recently developed by Horton[7] have been used.

Figure 4 is a polar plot of the received signal in db *versus* the angle about the edge of the semi-infinite screen for the case of an incident wave polarized with

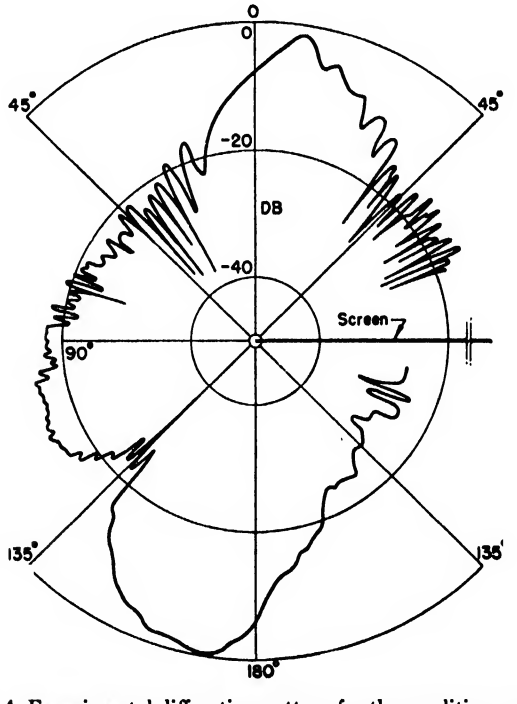


FIG. 4. Experimental diffraction pattern for the conditions of Fig. 2, except electric vector perpendicular to diffracting edge.

[6] See reference 2, p. 146.

[7] C. W. Horton, Phys. Rev. 75, 1263 (1949).

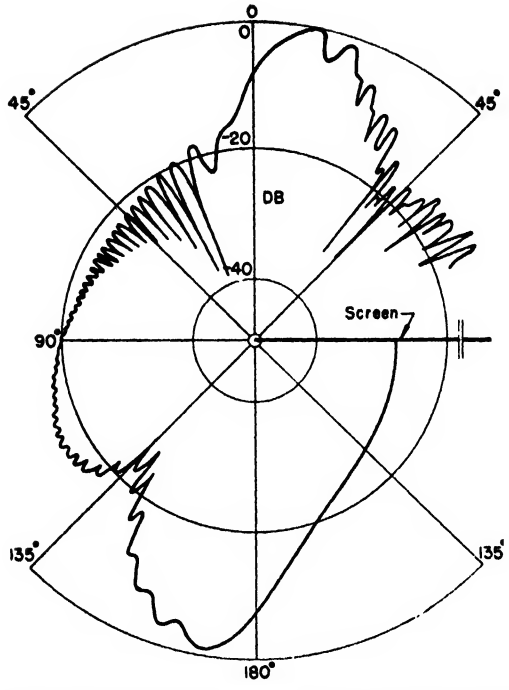


FIG. 5. Theoretical diffraction pattern for the conditions of Fig. 4.

the electric vector perpendicular to the diffracting edge of the screen. Again the horn was always turned so that it faced the edge of the screen. The theoretical pattern for the same conditions is shown in Fig. 5.

The process of resolving the total wave field into

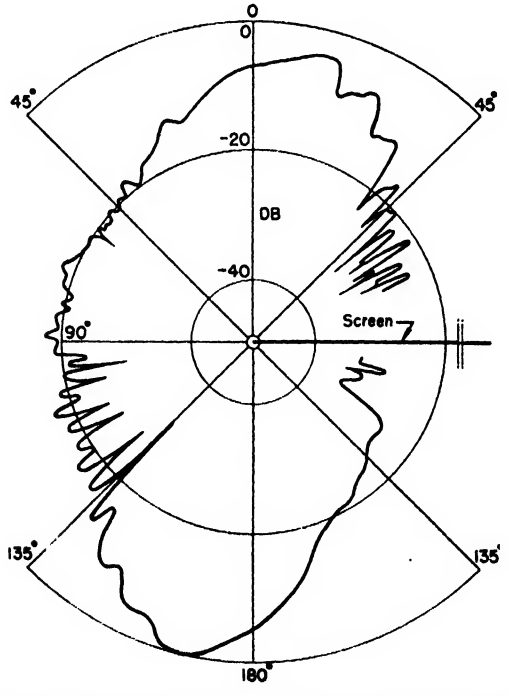


FIG. 6. Experimental diffraction pattern for semi-infinite screen for normal incidence, with incident wave from 0° direction. Receiving horn rotated in 15″ circle about center of diffracting edge. Electric vector parallel to diffracting edge.

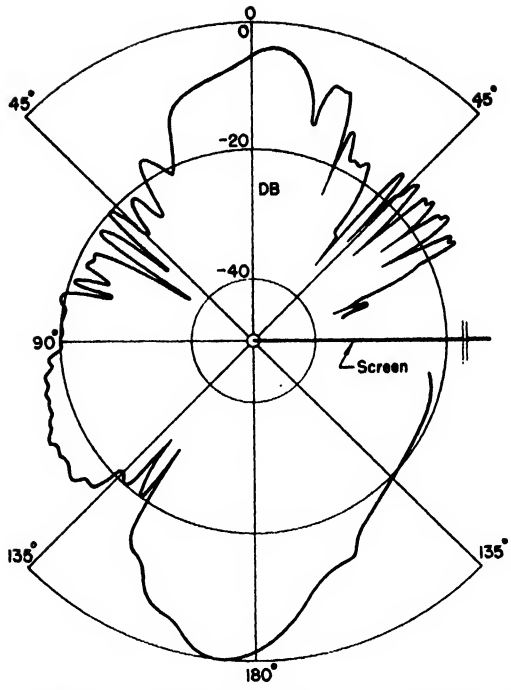


FIG. 7. Experimental diffraction pattern for the conditions of Fig. 6, except electric vector perpendicular to diffracting edge.

component fields has received a striking verification from these patterns. In the region of geometric reflection, the field can be resolved into three components; the incident plane wave, the reflected plane wave and a cylindrical wave diverging from the edge. To each of these the horn has a different response. Yet when the contributions of the three components are added, very good agreement with experimental patterns is obtained.

Figures 6 and 7 show the patterns for the case of the electric vector parallel and perpendicular respectively to the diffracting edge when the radius of the circle is 15 inches. The patterns were measured at the midpoint of the edge of the screen. The effective aperture angle of the horn is twice as large as in the case of the 30-inch circle and so the size of the lobes is correspondingly reduced.

Although an inspection of Figs. 2–5 is sufficient to show that there is good agreement between experiment and theory through all the angular range of the measurements it is of interest to compare for the shadow zone only the ratios of the amplitudes for the two cases of the electric vector parallel and perpendicular to the edge of the screen. Figure 8 shows the theoretical values of this ratio, the observed values for a circle of 15″ measured at the midpoint and at 6″ from the midpoint of the diffracting edge, and the values given by Maey[8] and Jentzsch[9] obtained by optical measurements. The

[8] E. Maey, Wied. Ann. 28, 117–130 (1886).

[9] F. Jentzsch, Ann. d. Phys. 84, 292–312 (1927). It might be mentioned that when Wolfsohn (reference 1, Fig. 6a, p. 275) cites Jentzsch's data, he titles the figure "intensity ratio" whereas the values plotted are "amplitude ratios."

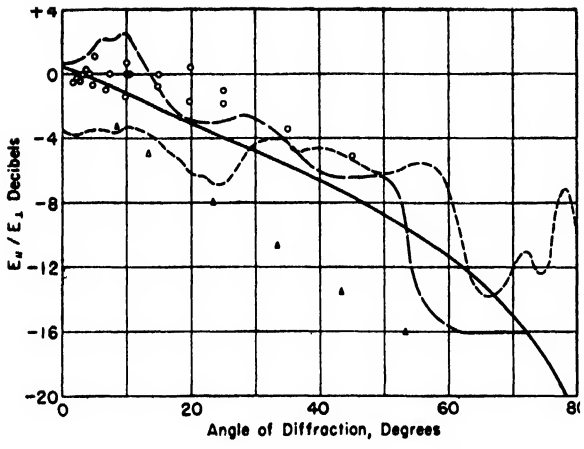


FIG. 8. Ratio of the amplitude of the measured electric vector for electric vector parallel to diffracting edge ($E_{\parallel}$) to that for electric vector perpendicular to diffracting edge ($E_\perp$) in the shadow zone behind the screen, *versus* diffraction angle. Normal incidence and semi-infinite screen. ——— Theory. - - - - Receiving horn rotated on 15″ circle, about center of diffracting edge. — — — Receiving horn rotated on 15″ circle, about point 6″ from center of diffracting edge. O, Optical data, Maey. Δ, Optical data, Jentzsch.

experimental data in Fig. 8 show clearly the effect of the interference fringes produced by the diffracted waves that come from the other three edges of the screen. It is difficult to draw a satisfactory average through the experimental curves. The general trend, however, is the same as the theoretical curve. Agreement is found with Maey's data, but Jentzsch's data

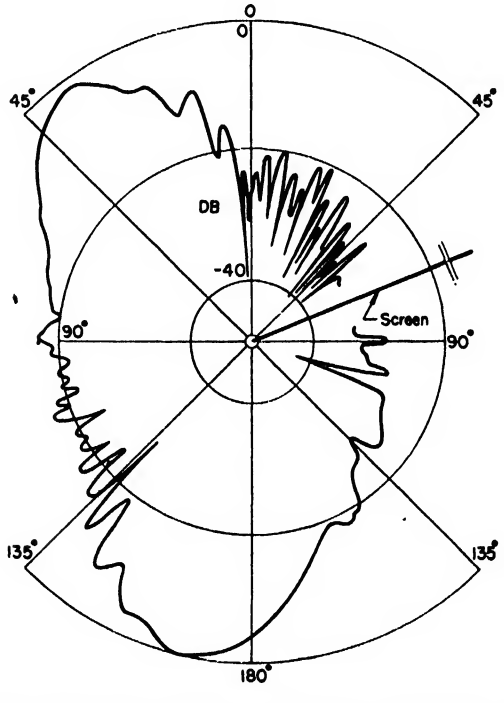


FIG. 9. Experimental diffraction pattern for semi-infinite screen for $-22\frac{1}{2}°$ incidence, with incident wave from 0° direction. Receiving horn rotated in 15″ circle about center of diffracting edge. Electric vector parallel to diffracting edge.

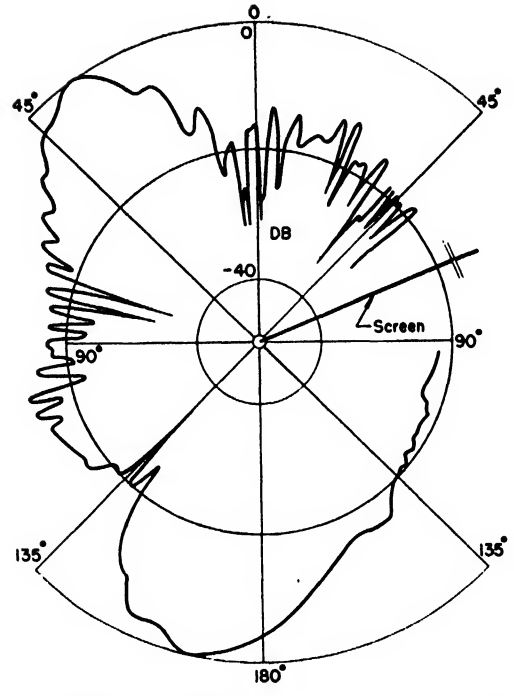


FIG. 10. Experimental diffraction pattern for the conditions of Fig. 9, except electric vector perpendicular to diffracting edge.

lie considerably below both the experimental and theoretical curves. At large angles the experimental data fall in general slightly above the theory. This effect may be in part explained by the smallness of the component values used in calculating the ratios. For such small values, inherent noise in the measuring apparatus plus diffracted contributions from surrounding objects as well as the other three edges of the screen may well increase the smaller component value more than the larger, and thus increase the ratio. Similar curves of $E_{\parallel}/E_\perp$ are obtained for the measurements at a radius of 30″ but the errors of measurement are larger because of the increased diffraction about the other three edges.

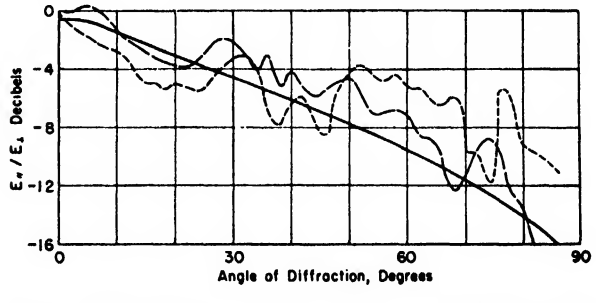


FIG. 11. Ratio of the amplitude of the measured electric vector for electric vector parallel to diffracting edge to that for electric vector perpendicular to diffracting edge in the shadow zone behind the screen, *versus* diffraction angle. Angle of incidence $= -22\frac{1}{2}°$; semi-infinite screen. ——— Theory. - - - - Receiving horn rotated on 15″ circle, about center of diffracting edge. — — — Receiving horn rotated on 15″ circle, about point 6″ from center of diffracting edge.

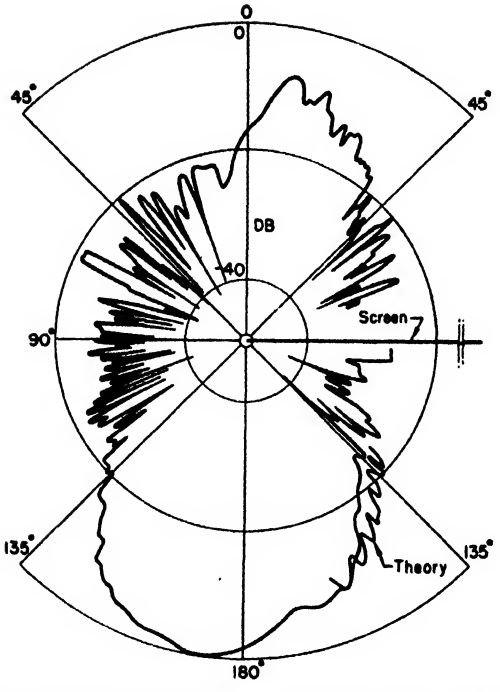


FIG. 12. Experimental and theoretical diffraction patterns for quarter-infinite screen for normal incidence, with incident wave from 0° direction. Receiving horn rotated in 30″ circle about top diffracting edge. Electric vector parallel to vertical diffracting edge.

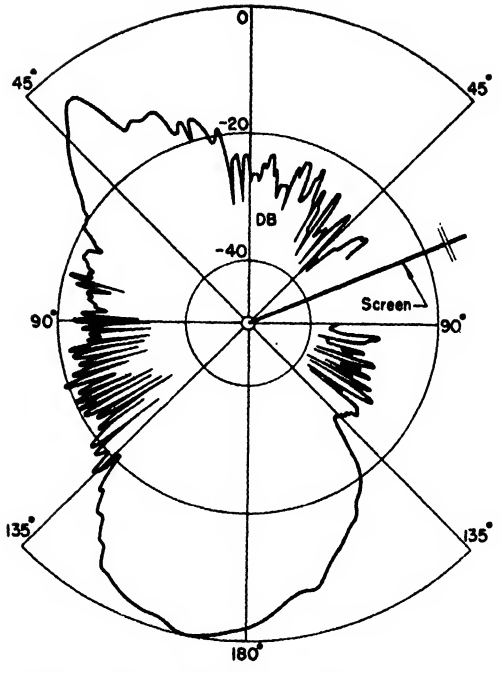


FIG. 14. Experimental diffraction pattern for quarter-infinite screen for $-22\frac{1}{2}°$ incidence, with incident wave from 0° direction. Receiving horn rotated in 30″ circle about top diffracting edge. Electric vector parallel to vertical diffracting edge.

IV. ANGLE OF INCIDENCE −22.5 DEGREES

The measurements described above were repeated for the case that the direction of propagation of the incident wave made an angle of −22.5 degrees with the normal to the screen. Figures 9 and 10 show the two patterns for the cases of the electric vector parallel

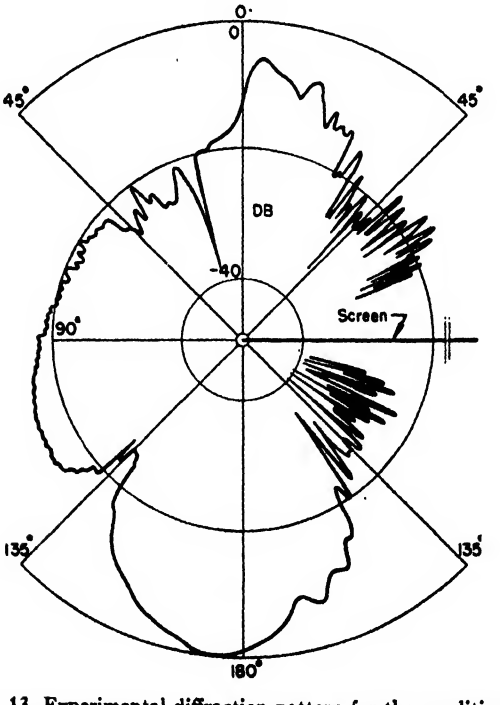


FIG. 13. Experimental diffraction pattern for the conditions of Fig. 12, except electric vector perpendicular to vertical diffracting edge.

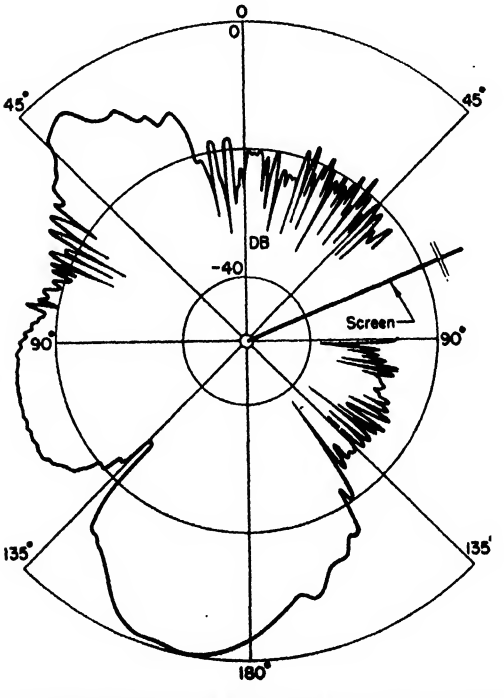


FIG. 15. Experimental diffraction pattern for the conditions of Fig. 14, except electric vector perpendicular to vertical diffracting edge.

and perpendicular to the diffracting edge of the screen respectively. Figure 11 shows a comparison between the theoretical and experimental values of the ratio of the amplitudes for the electric vector parallel and perpendicular to the edge of the screen. The experimental curves in Fig. 11 follow the theory quite well, with some deviations due to effects of extraneous diffraction. The experimental points tend to rise slightly above the theory at large angles, due in part to incidental noise and diffraction effects from the other edges.

V. QUARTER-INFINITE SCREEN

A rigorous theoretical solution of the diffraction pattern of a quarter-infinite screen has not been obtained. A first approximation is obtained by assuming both edges of the quarter screen to be infinite in extent so that the analysis described above may be applied to each edge. The effect of termination of the edges at the point of intersection is recognized by assuming a source of spherical waves at the corner. Contributions from the two diffracting edges, the incident wave, and the point at the corner are then added together to give the solution. Detailed calculations were carried out for the case of the electric vector parallel to the vertical diffracting edge, when the receiving horn is rotated in a circle whose plane contains the upper edge of the screen. These calculations were made for normal incidence only, and were restricted to the region in the shadow zone in which the angle of diffraction with respect to the vertical edge was between 20° and 50°. This restriction was made since it was felt that such an approximate analysis should apply best in the shadow region behind the screen.

Figures 12 and 13 show the experimental measurements, which were carried out for all angles for the case of normal incidence on the quarter-infinite screen. The theoretical data are plotted in the lower right quadrant of Fig. 12. The experimental and theoretical curves are much alike, but the theory shows larger lobes and a somewhat different slope from the experiment. A comparison of Figs. 12 and 13 with Figs. 2 and 4, respectively, shows that the diffraction pattern of a quarter-infinite screen is surprisingly similar to that of a semi-infinite screen.

Figures 14 and 15 show the experimental measurements for a quarter-infinite screen when the normal to the screen makes an angle of $-22.5°$ with the direction of the incident wave. Further similarity to patterns for the semi-infinite screen is observed by comparison with Figs. 9 and 10.

Measurement of the Surface Temperature of Evaporating Water Drops*

JOHN C. JOHNSON
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received July 14, 1949)

Measurements made of the surface temperature of evaporating water drops are in agreement with the requirements of the theory ennunciated by Fuchs which states that the temperature is determined by the difference between the ambient and saturated vapor density of the drop.

INTRODUCTION AND THEORY

FOR some years, studies, both theoretical and experimental, have been conducted on the rate of evaporation of liquids in the form of spherical drops. Studies using water droplets have been of particular interest to meteorologists, but evaporation processes are of concern in studies of liquid smokes, in atomizing fuel oils, and in studies of other volatile liquids. Both theory and experiment show that the drop diameter varies as a parabolic function of time when evaporation (or condensation) occurs, so that the surface area changes directly with the lapse in time. This leads to an expression of the type (Houghton 1933)

$$a^2=a_0^2-(8k/\bar{\rho})(\rho_0-\rho)t \tag{1}$$

obtained from the differential equation

$$dM/dt=-2\pi ka(\rho_0-\rho), \tag{2}$$

where dM/dt = the rate of change of the mass of the liquid, k = diffusivity of the vapor in the atmosphere surrounding the drop, a = diameter of the drop at time t, a_0 = diameter at $t=0$, $\bar{\rho}$ = density of the liquid, and $(\rho_0-\rho)$ is the difference in vapor density at the surface of the drop and the surroundings, the latter ambient condition being essentially the concentration of the vapor in the atmosphere in the absence of or at infinite distance from the drop.

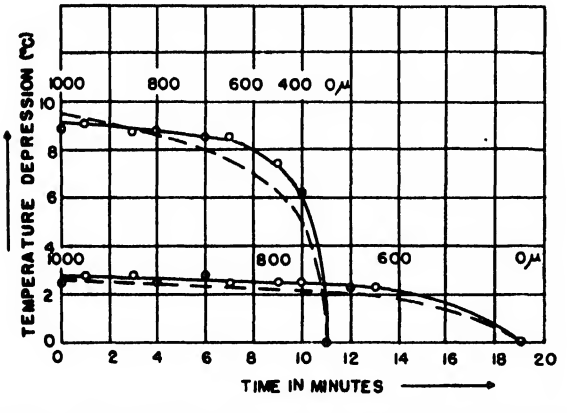


FIG. 1. The solid line of the upper curve is the measured data for a drop evaporating in an atmosphere with temperature and relative humidity of 25.4°C and 37 percent respectively. Number 2 thermocouple was used. The lower solid curve shows the experimental results using thermocouple No. 4, where the ambient temperature and relative humidity were 23.9°C and 68 percent respectively. The dashed curves are a plot of Eq. (6) using the same values of $(\rho-\rho_0)$ that were measured experimentally. $C=2.1\times10^{-4}$ was used.

In practice, the use of Eq. (1) requires a knowledge of the quantity $(\rho_0-\rho)$. Confining the discussion for the moment to water vapor, ρ is the ambient vapor density evaluated from the measurement of the relative humidity of the air, the temperature of the air, and a set of vapor pressure tables for water (neglecting the slight effect on this pressure by the pressure of a neutral gas). The quantity ρ_0, is the saturation vapor density at the temperature of the surface of the evaporating drop. Reason dictates that this latter temperature be colder than the surrounding atmosphere for an evaporating drop. The wet bulb temperature has been suggested for this temperature, but does not give the correct result. Fuchs (1934) [see Houghton (1938)] attacked this problem by considering the energy lost to the drop by diffusion to be balanced in the steady state by the conduction of heat from the air to the surface of the drop. If Eq. (2) be multiplied by the latent heat of vaporization, L, and equated to the expression for the radial conduction of heat by the air to the drop

$$dQ/dt=-2\pi K(T-T_0)a \tag{3}$$

where K is the thermal conductivity of the air, there is obtained a thermal condition

$$T-T_0=(kL/K)(\rho_0-\rho) \tag{4}$$

which must be satisfied as evaporation (or condensation) occurs. T and ρ are the ambient temperature and vapor density, while T_0 and ρ_0 refer to saturation conditions at the drop surface. As $\rho_0=f(T_0)$, Eq. (4) leads to a "cut and trial" solution. It is the purpose of this paper to present an experimental verification of Eq. (4), by direct measurement of the surface temperatures of evaporating water drops.

DISCUSSION OF EXPERIMENTAL PROCEDURE AND CORRECTIONS

The drop under consideration can be suspended from a small thermocouple in a suitable chamber where the humidity and temperature of the chamber's atmosphere can be controlled and measured. If there were no thermal losses due to heat from the air being conducted to the thermocouple junction by the thermocouple wires

* Work done in connection with the partial fulfillment of the degree Sc.D. at Massachusetts Institute of Technology, July 1948.

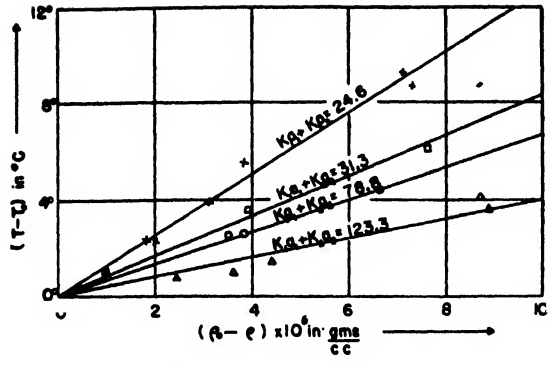


FIG. 2. Data used in determining the best value of A for the four different thermocouples. A is the slope of the curve, $(T-T_0)/(\rho_0-\rho)$.

TABLE I. Thermocouple data.

Thermocouple	No. 1	No. 2	No. 3	No. 4	Units
a_1	127	24	28	79	microns
a_2	127	49	104	117	microns
K_1	0.918	0.918	0.918	0.918	cal./cm°C
K_2	0.054	0.054	0.054	0.054	cal./cm°C
$K_1a_1+K_2a_2$	123.3	24.6	31.3	78.8	cal. microns / cm °C

$$\left(\frac{kL}{K}\right)_{20°C}=2.45\times10^{6}\frac{°C}{g}cc$$

themselves, and thereby helping to compensate the cooling effect of the evaporating drop on the thermocouple junction, the ratio $(T-T_0)/(\rho_0-\rho)$ could be directly measured and compared at once with the constant kL/K predicted by theory. Due to the error introduced by any real thermocouple, it is necessary to use several thermocouples with different heat losses for the measurements, extrapolating the results to one of zero loss for the comparison. If the two lead wires to the thermocouple junction are considered to be cylinders of uniform diameters a_1 and a_2, with conductivities K_1 and K_2 respectively, the heat loss from the wires is to the first approximation (Johnson 1948)

$$(dQ/dt)_{\text{wire}}=-2\pi C(K_1a_1+K_2a_2)(T-T_0), \qquad (5)$$

where C is a dimensionless constant due to the geometry of the system having an empirical value of $2.5\pm0.4\times10^{-3}$. The loss of energy by diffusion (Eq. (2) multiplied by the latent heat of vaporization of the liquid) is just balanced by the conduction of heat to the thermocouple junction both from the surface of the drop and from the wire to give

$$T-T_0=\frac{kL(\rho-\rho_0)}{K\{1+C[(K_1a_1+K_2a_2)/Ka]\}}=A(\rho-\rho_0). \qquad (6)$$

An inspection of Eq. (6) shows that A is a function of the drop diameter for a given thermocouple. For large drops or very fine wires $A\rightarrow kL/K$ and the second term in the denominator is small. At the other limit where the drop is very small in relation to the wires, $A\rightarrow0$ and no difference in temperature is measured. There is an

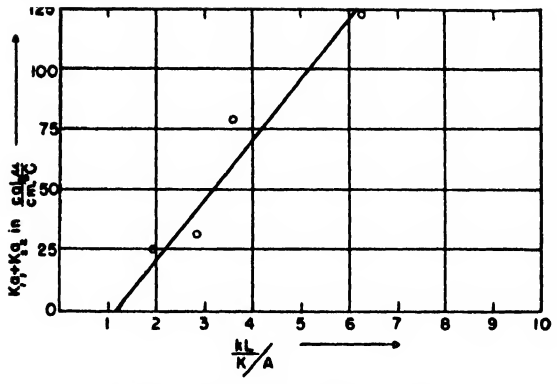


FIG. 3. An extrapolation of the observed data to a thermocouple having no thermal influence on the evaporating drop, to approximate the conditions of a cloud or small rain drop evaporating in the free atmosphere (assuming zero fall velocity).

intermediary region where $a\cong C(K_1a_1+K_2a_2/K)$ causing A to change rapidly. An example of this is given in Fig. 1. The temperature changes for the evaporating drop are shown for two constant vapor density differences.

RESULTS AND CONCLUSIONS

By observing the evaporation of several drops from a given thermocouple, the best value of A can be determined for the thermocouple. This was done for four thermocouples at a drop diameter where the temperature difference was large and nearly independent of drop diameter. The results are given in Fig. 2. Finally the ratio $(kL/K)/A$ is plotted against the wire constant $K_1a_1+K_2a_2$ and extrapolated to zero. Equation (6) on rearranging to

$$(kL/K)/A=(C/Ka)(K_1a_1+K_2a_2)+1 \qquad (7)$$

indicates that the relationship is linear when all measurements are made at one value of the drop diameter. This latter condition has been substantially met in Fig. 3. Least square analysis indicates that

$$(kL/K)/A_0=1.15 \quad A\rightarrow A_0 \text{ as } K_1a_1+K_2a_2\rightarrow0$$

for this series of measurements. From the experimental evidence, it appears that Eq. (4) adequately describes the surface temperature of an evaporation drop of water under steady state conditions where extreme conditions of radiation and convection are lacking. The values of the wire constants are given in Table I.

BIBLIOGRAPHY

(1) N. Fuchs, "Uber die Verdampfungsgeschwindigkeit kleiner Tröpfchen in einer Gasatomsphäre," Physik. Zeits. d. Sowjetunion 6, 224–243.
(2) H. G. Houghton, Physics 4, 419 (1933).
(3) H. G. Houghton and W. H. Radford, "On the Local Dissipation of Natural Fog," Papers in Physical Oceanography and Meteorology, Mass. Inst. of Technology and Woods Hole Oceanographic Institute 6, No. 3, pp. 1–63 (October 1938).
(4) J. C. Johnson, "A Study of the Freezing Temperature of Supercooled Water with Special Reference to Drops Having Diameters from 150 to 1700 Microns," Sc.D. thesis, M.I.T. (July 1948).

X-Ray Measurement of Order in Single Crystals of Cu_3Au*,**

J. M. Cowley†
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received August 4, 1949)

X-ray diffraction methods have been used to measure long- and short-range order parameters for single crystals of the alloy Cu_3Au held at elevated temperatures. It is shown that short-range order parameters may be obtained by a three-dimensional Fourier analysis of the "scattering power" for the diffuse background scattering of x-rays, expressed as a function of reciprocal lattice coordinates. Appropriate experimental and computational procedures are outlined. Intensity measurements have been made with a special arrangement of a bent-crystal monochromator and a Geiger-counter spectrometer. Short-range parameters are given for the first ten shells of atoms surrounding a given atom for three temperatures above the critical temperature of Cu_3Au. These results, and those for the long-range order parameter agree well with theoretical predictions.

INTRODUCTION

X-RAY diffraction methods allow quantitative determinations to be made of the amount of order present in an alloy. When long-range order is present, super-lattice reflections appear in the x-ray diffraction pattern with intensity proportional to the square of S, the long-range order parameter defined by Bragg and Williams.[1] Above the critical temperature of ordering of the alloy, the long-range order is zero, but short-range order persists in that there is a tendency for each atom to be surrounded by more than an average number of unlike atoms. In this case, the randomness in the lattice gives rise to a diffuse background scattering of x-rays, and the presence of short-range order gives broad maxima in this diffuse background at the positions of the super-lattice reflections of the ordered lattice. Values of the parameters defining short-range order may be obtained by an analysis of the diffuse scattering.

Measurements of short-range order parameters have been made by Wilchinsky[2] using powder samples of the alloy Cu_3Au. The accuracy of this work was limited by the difficulty of making an accurate separation of the diffuse scattering due to the thermal motion of the atoms from that due to the state of short-range order. This difficulty is largely overcome by the use of a single crystal. In a single-crystal diffraction pattern, the thermal diffuse scattering occurs mostly about the strong "fundamental" reflections, well separated from the maxima of the diffuse scattering due to short-range order. Hence the two effects may be distinguished.

Measurements of long-range order in Cu_3Au have been made by Wilchinsky[2] and by Komar and Buinov,[3] using powder samples. Chipman and Warren[4] have measured the long-range order in a single crystal of the alloy CuZn. The single crystal method has the advantage of giving greater intensity for Geiger-counter measurements, and also permits the detailed study of the time-variation of intensity and shape of super-lattice reflections.

MEASUREMENT OF SHORT-RANGE ORDER IN CU_3AU

Diffraction Theory for Short-Range Order

When an x-ray beam is reflected from a crystal, the intensity reflected in a given direction may be expressed in electron units, in the form

$$I=\sum_m \sum_n f_m f_n \exp\left[\frac{2\pi i}{\lambda}(\bar{S}-\bar{S}_0)\cdot(\bar{r}_m-\bar{r}_n)\right],$$

where f_m and f_n are the atomic scattering factors for the atoms at the end points of the vectors $\bar{r}_m$ and $\bar{r}_n$ drawn from some origin, and $\bar{S}_0$ and $\bar{S}$ are the unit vectors representing the incident and reflected beams. The summations are made over all atoms of the crystal.

We consider a crystal composed of N unit cells of simple structure, such as body-centered or face-centered cubic, with A and B atoms present in the proportions m_A and m_B. If short-range order is present in the crystal, there will be a relationship between the value of f_m and the values of the f_n for atoms close to the atom denoted by m. We say that among the c_i atoms which are i^{th} nearest neighbors of a B atom, on the average n_i are A atoms. If, then, such relationships are incorporated in the expression for the intensity, the scattered intensity is found to consist of sharp "crystalline" reflections plus a diffuse background intensity given by

$$I_D=Nm_Am_B(f_A-f_B)^2 \times\left\{1+\sum_{i=1}^{\infty}\sum_{j=1}^{c_i}\left(1-\frac{n_i}{m_Ac_i}\right)\exp\left[\frac{2\pi i}{\lambda}(\bar{S}-\bar{S}_0)\cdot\bar{r}_{ij}\right]\right\},$$

where $\bar{r}_{ij}$ is the vector from the origin to the j^{th} atom

* Extract from a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics at the Massachusetts Institute of Technology.

** Work supported in part by the ONR.

† Now with the Division of Industrial Chemistry, Commonwealth Scientific and Industrial Research Organization, Australia.

[1] Bragg and Williams, Proc. Roy. Soc. A145, 699 (1934).
[2] Z. W. Wilchinsky, J. App. Phys. 15, 806 (1944).
[3] Komar and Buinov, J. Phys. (USSR) 11, No. 5 (1947).
[4] B. E. Warren and D. Chipman, Phys. Rev. 75, 1629 (1949).

of the i^{th} shell of atoms surrounding a B atom. The vector $\bar{S}-\bar{S}_0$ may be expressed in terms of the continuous coordinates h_1, h_2, h_3, and the vectors $\bar{b}_1$, $\bar{b}_2$, $\bar{b}_3$, of the reciprocal lattice as

$$\bar{S}-\bar{S}_0=(h_1\bar{b}_1+h_2\bar{b}_2+h_3\bar{b}_3)\lambda.$$

Similarly the vector $\bar{r}_{ij}$ may be expressed in terms of the coordinates l,m,n, of the atoms, and the vectors $\bar{a}_1$, $\bar{a}_2$, $\bar{a}_3$, of the crystal lattice, as $\bar{r}_{ij}=l\bar{a}_1+m\bar{a}_2+n\bar{a}_3$. Putting $\alpha_{lmn}=\alpha_i=l-(n_i/m_AC_i)$, the summation in the intensity expression, abbreviated by I_D', then becomes

$$I_D'=\sum_{l=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty}\alpha_{lmn}\exp[-2\pi i(lh_1+mh_2+nh_3)].$$

Thus I_D' is expressed by a three-dimensional Fourier series in terms of the reciprocal lattice coordinates, and is a periodic function with the periodicity of the reciprocal-lattice unit cell.

The coefficients of the Fourier series are given in the usual way as

$$\alpha_{lmn}=\int_0^1\int_0^1\int_0^1 I_D'\exp[2\pi i(lh_1+mh_2+nh_3)]dh_1dh_2dh_3.$$

From the diffraction theory, these Fourier coefficients are given as $\alpha_{lmn}=\alpha_i=l-(n_i/m_AC_i)$; or, in terms of the probability p_{lmn} that the atom with coordinates l,m,n with respect to a B atom should be an A atom, $\alpha_{lmn}=1-p_{lmn}/m_A$.

The probability that an atom is an A atom is given by $P_{lmn}=m_A(1-\alpha_{lmn})$ if a B atom is at the origin, and by $P_{lmn}=m_A+m_B\alpha_{lmn}$ if the origin is an A atom.

The Fourier coefficients have the conventional features of short-range order parameters in that they are zero for the completely random arrangement, and have a maximum (absolute) value for the perfectly ordered state. For even-numbered shells the value for perfect order is unity. For odd-numbered shells the value for perfect order is unity. For odd-numbered shells, the value for perfect order is negative and may be fractional. For example, for Cu_3Au type alloys, α_i°, the value of α_i for perfect order, is equal to $-1/3$, for i odd. It is evident that short-range order parameters defined in this way are the best suited to considerations of the x-ray diffraction effects. It has also been found that they are very convenient for theoretical considerations of order in alloys.[5] Hence the coefficients α_{lmn} or α_i have been used as short-range order parameters in preference to parameters defined in analogy with Bethe's[6] σ. For AB alloys, such as CuZn, the α_{lmn} are equivalent to the r_{hkl} defined by Zernike,[7] and for this case, $\alpha_1=-\sigma$.

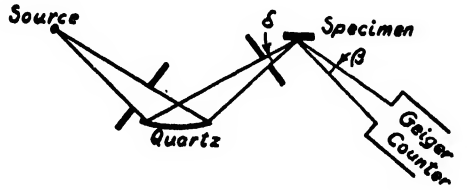


FIG. 1. Arrangement used for measurements of diffuse scattering from single crystals, using a bent-crystal monochromator.

It may be shown readily[6] that the limiting value of α_i for i very large is given by $\alpha_i^\circ\cdot S^2$. Thus the value of α_i falls off to zero with increasing distance from the origin above the critical temperature, but approaches a finite limit when long-range order is present. For the case of finite long-range order, the Fourier series representing the scattering power in reciprocal lattice space may be divided into two series, as MacGillavry and Strijk[8] have suggested: one series with coefficients equal to $\alpha_i^\circ\cdot S^2$, representing the sharp super-lattice reflections with integrated intensity proportional to S^2, and the other with coefficients $\alpha_i-\alpha_i^\circ\cdot S^2$, giving broad maxima in the diffuse background surrounding the sharp super-lattice reflections.

The experimental determination of short-range order parameters involves the measurement of a large number of intensities of the diffuse background scattering of x-rays from a single crystal, the calculation of the quantity I_D' as a function of the reciprocal lattice coordinates, and the three-dimensional Fourier analysis of this function.

Experimental Methods

Measurements were made of the intensity of reflection of copper $K\alpha$ radiation from the face of a single crystal of Cu_3Au cut parallel to the (1,0,0) lattice plane. The crystal was held at a temperature above the critical temperature of Cu_3Au, and intensity measurements were made with a conventional Geiger counter and scaling circuit.

For measurements of diffuse background intensities it is necessary to use monochromatic radiation such as that given by a crystal monochromator. However, it was found that the intensity given by a plane monochromator, such as a well-ground rocksalt cube face,

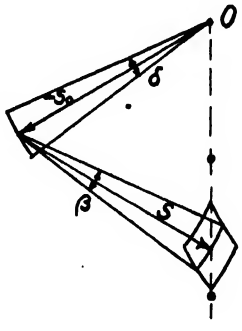


FIG. 2. The volume-element of reciprocal lattice space defined by the divergence of the x-ray beams.

[5] J. M. Cowley, "An approximate theory of order in alloys" (submitted for publication in Phys. Rev.).
[6] H. A. Bethe, Proc. Roy. Soc. A150, 552 (1935).
[7] F. Zernike, Physica 7, 565 (1940).
[8] C. H. MacGillavry and B. Strikj, Physica 11, 369 (1946).

was insufficient to allow accurate intensity measurements with a Geiger counter. Hence a focusing monochromator, using an elastically bent quartz crystal, was employed in an arrangement sketched in Fig. 1. This arrangement gives sufficient intensity by taking advantage of the fact that a low divergence beam, such as that given by a plane crystal monochromator, is not necessary for the measurement of slowly-varying background intensities, but a relatively large divergence of incident and reflected beams may be used without serious decrease in the accuracy of measurement. In

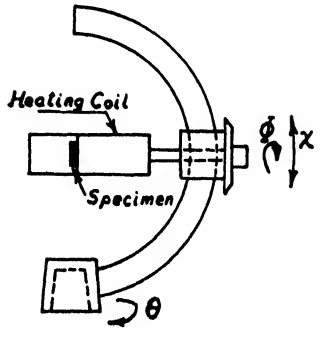


FIG. 3. Specimen mount, for measurement of intensities reflected from single crystal faces at elevated temperatures.

Fig. 1, the incident beam has an effective divergence δ in the horizontal plane. The rectangular aperture at the entrance to the Geiger counter gives the reflected beam a divergence β in the same plane and a divergence γ, not shown, in the perpendicular plane. Figure 2 shows the equivalent of this arrangement in the reciprocal lattice. The vector, S_0, representing the incident beam, varies through an angle δ. The vector S representing the reflected beam varies through an angle β in the same plane and an angle γ in the perpendicular plane. All the "scattering power" contained in the volume element of reciprocal space so defined contributes to the measured intensity. In practice the angles δ, β and γ were made equal to three degrees.

In order to obtain measurements of the intensity for orientations of the single crystal corresponding to various positions of the volume element in reciprocal lattice space, the specimen mount sketched in Fig. 3 was used. The large semicircular arcs stood in a vertical plane on the table of a spectrometer. The specimen was heated by a heating coil and surrounded by a shield having nickel or beryllium windows. An atmosphere of hydrogen was maintained within the shield to prevent oxidation of the surface of the crystal. Three calibrated rotations are provided; rotation through the Bragg angle θ, rotation about the (1,0,0) crystal axis, perpendicular to the crystal face, through the angle Φ, and a tilting of this axis through an angle χ in a vertical plane.

The function I_D' may be found throughout the whole of a unit cell of the reciprocal lattice by applying symmetry operations to values found for a relatively small portion of the unit cell. Thus, for Cu_3Au, intensity measurements were made, and values of I_D' were calculated only for the region contained in the triangular prism *ABCDEF* shown in Fig. 4. This prism has one thirty-second of the volume of the whole unit cell, which may be built up from the prism by use of the center of symmetry, *G*, and planes of symmetry.

From the observed intensities, the values of I_D' were calculated using the relation

$$I_D' = K(I_{Obs} - I_T - I_C)\left[\frac{(\sin 2\theta)}{(1+\cos^2 2\theta)}\right]\frac{1}{(f_{Au}-f_{Cu})^2}\frac{1}{A(\theta)},$$

where I_T and I_C represent the "temperature-diffuse" and Compton scattering contributions to the diffuse background intensity, K is a constant, and $A(\theta)$ is a correction for the absorption in the nickel or beryllium foil surrounding the specimen. Since the experimental conditions were such that the angles made by the incident and reflected beams with the crystal face were equal, no correction for absorption in the specimen was necessary.

The contribution of the Compton scattering was calculated to be approximately one-tenth of the average value of I_D'. The temperature diffuse scattering was evaluated by measuring diffuse background intensities at temperatures of 360°C, which is just below the critical temperature of about 385°C, and room temperature. At 360°C, the diffuse scattering consisted of temperature-diffuse scattering, Compton scattering, and a small amount of scattering due to disorder. For a specimen cooled quickly from 360°C to room temperature, the amounts of Compton and disorder scattering were unchanged, but the temperature-diffuse scattering was decreased. Thus the rate of variation of the temperature-diffuse scattering was obtained, and by extrapolation to absolute zero of temperature, the amount of disorder scattering at 360°C was found. The correction term for temperature-diffuse and Compton scattering, to be subtracted from the observed intensity at any elevated temperature, was then defined as the intensity measured at 360°C, minus the contribution of the disorder scattering at 360°C, plus a term proportional to the difference between the temperature of observation and 360°C.

At each of three temperatures greater than the critical temperature, more than 450 intensity measurements were made for settings of the angles θ, Φ, and χ, corresponding to positions of the measuring volume-element spaced regularly throughout the portion of the reciprocal lattice cell shown in Fig. 4. The value of I_D' was calculated for each position. Values of I_D' were then interpolated on to a cubic net of points spaced at intervals of 0.05 in the coordinates h_1, h_2, h_3, which are so defined that the unit of distance is the distance of the (2,0,0) reflection from the origin of the reciprocal lattice. By use of appropriate symmetry operations, the values of I_D' at the 8000 points of this cubic net within the unit cell were obtained.

The values of the Fourier coefficients α_{lmn} are given

by the integration of the product of I_D' and the appropriate exponential function throughout the unit cell of the reciprocal lattice. In practice, the presence of a center of symmetry allows the exponential functions to be replaced by cosines, and the integration is approximated by a summation of the values of the product at regularly spaced points. Thus,

$$\alpha_{lmn}=K\sum_{h_1}\sum_{h_2}\sum_{h_3} I_D' \cos 2\pi(lh_1+mh_2+nh_3).$$

The summation was carried out over the 4000 points of the cubic net contained in half of the unit cell. The normalizing constant, k, was obtained from the relation,

$$\alpha_{000}=K\sum_{h_1}\sum_{h_2}\sum_{h_3} I_D'.$$

The summation was greatly simplified when one or two of the atom coordinates were zero. Thus

$$\alpha_{l00}=K\sum_{h_1}(\sum_{h_2}\sum_{h_3} I_D') \cos 2\pi l h_1$$

and

$$\alpha_{lm0}=K\sum_{h_1}\sum_{h_2}(\sum_{h_3} I_D') \cos 2\pi(lh_1+mh_2).$$

For the general case, in which no such simplification was possible, a "template" method of performing the multiplication and addition was used. The values of I_D' were tabulated at regularly spaced positions for values of h_2 and h_3 from 0 to 1, for each value of h_1. Templates were made by cutting holes in stiff paper so that, when the templates were placed over the tables of values of I_D', only numbers to be multiplied by given values of the cosine functions were visible.

In some cases the values of the α_{lmn} so obtained did not show symmetry with respect to the three reciprocal lattice aces. Thus, for example, the values obtained for α_{200} and α_{020} were different. Such differences arose from imperfect correction for temperature-diffuse scattering, and were small for carefully determined backgrounds. In order to minimize the errors arising from such asymmetry, the values of the α_{lmn} were obtained for various permutations of the indices, and the weighted average was found.

Experimental Results for Short-Range Order

Intensity measurements were made for Cu_3Au at temperatures of 405, 460, and 550°C. The distribution of scattering power in the reciprocal lattice, calculated from these measurements, showed interesting features. Figure 5 is a contour plot of the scattering power due to short-range order, in arbitrary units, in the (1,0,0) plane of the reciprocal lattice of Cu_3Au at 405°C. The corners of the square are sites of the sharp fundamental reflections. The diffuse maxima at the super-lattice reflection positions appear to be flattened into a disk-like shape. At the center of the square, one of these disks is seen in section by the plane of the disk, and on each side of the square is the section of a disk with its plane perpendicular to the plane of the section. It is seen that the flattened sides of the disks face the nearest fundamental reflections, and each fundamental reflection is surrounded by six such flattened sides.

The disk-like shape of the diffuse maxima seems to be a consequence of the structure of the Cu_3Au lattice, rather than any particular values for the short-range order parameters. Qualitatively, the same shape is given by calculation assuming a wide variety of values for the order parameters for the first two shells, ignoring the effects of further parameters.

The same disk-like shape has been found, on a much smaller scale, for the relatively sharp super-lattice reflections from single crystals of Cu_3Au[9,10] for temperatures below the critical temperature. For this case, the shape of the super-lattice reflections was explained by Wilson[11] in terms of the domain-structure of the ordered alloy.

The short-range order parameters calculated by Fourier analysis of such distributions of scattering-power are given in Table I. The third column of the table gives the coordinates of representative points of the successive shells, referred to half of the length of the face-centered cubic unit cell as unit distance. The values of the order parameters for a perfectly ordered lattice are given in the fourth column. The magnitudes of the short-range order parameters decrease, in general, with increasing temperature and with increasing distance from the origin. At 405°C, the magnitude of the nearest-neighbor order parameter is seen to be almost half that for perfect order. The

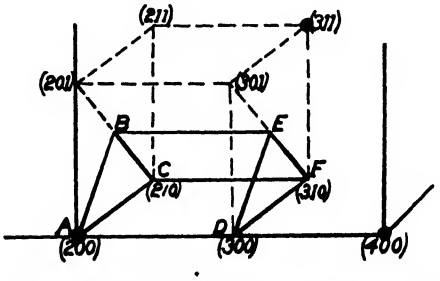


FIG. 4. The prism-shaped portion of reciprocal-lattice space, $ABCDEF$, throughout which the scattering power was determined.

parameters for the fifth and tenth shells indicate that, with a gold atom at the center, there is an excess of one copper atom among the 24 atoms of the fifth shell, and a defect of half a copper atom among the 24 atoms of the tenth shell. Thus the influence of the atom at the origin is appreciable at greater distances than those which have been considered in previous theoretical and experimental work.

[9] B. Strijk and C. H. MacGillavry, Physica **12**, 129 (1946).
[10] Edmunds, Hinde, and Lipson, Nature **160**, 304 (1947).
[11] A. J. C. Wilson, Proc. Roy. Soc. **A181** (1943), Nature **160**, 305 (1947).

It is difficult to make an accurate estimate of the probable error in the values of the short-range order parameters found. However, it is believed that the error in the parameters for the inner shells does not exceed 0.01. For more distant shells, the absolute error is probably slightly less, and the relative error greater. The principal sources of error are inaccuracies in the correction for temperature-diffuse scattering and in the interpolation of the values of I_D' from a conical-polar to a cubic net of points.

In general, the experimental order parameters given

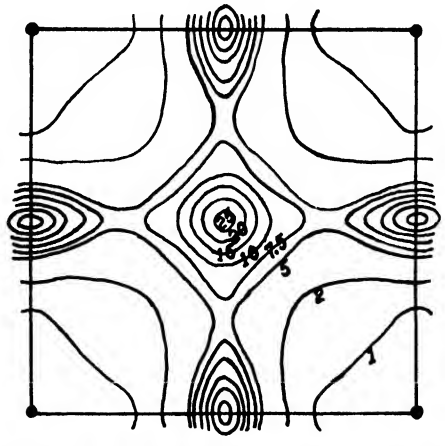


FIG. 5. Contour plot of scattering power due to short-range order in the (1,0,0) plane of the reciprocal lattice of Cu_3Au at 405°C.

in Table I have the same sign as the corresponding parameters for the case of perfect order. However, for some of the parameters for the third, sixth, and ninth shells, the signs are reversed, indicating a defect of copper atoms in a shell surrounding a gold atom instead of an excess, or an excess instead of a defect. The significance of these reversals of sign becomes apparent when the number of copper atoms in excess or defect in each shell is considered as a function of the radius of the shell. In Fig. 6 the number of copper atoms in defect (as compared with the number for a completely random arrangement) for each shell, given by the product $\frac{3}{4}c_i\alpha_i$, is plotted as a vertical line at the abscissa value corresponding to the radius of the shell. This is done for perfect order and for temperatures of 405 and 550°C. For the case of perfect order, alternate shells show an excess and a defect. However, for the elevated temperatures there is a marked tendency for groups of shells to show a common excess or defect, so that the alternation is less rapid. Thus, after the large excess of copper atoms in the first shell about a gold atom, the next three shells have a defect, the next three have a common tendency to show an excess, and so on.

One way of interpreting this behavior is to consider that the predominating force is that tending to surround each atom with unlike neighbors. The first shell surrounding a gold atom is thus copper-rich. All atomic sites which are nearest neighbors to atoms in this first shell will then tend to be gold-rich. The atoms of the second, third, and fourth shells are all nearest neighbors of points in the first shell, and so will show a common defect in copper atoms. Similarly, the atoms of the fifth, sixth, and seventh shells are nearest neighbors of atoms in these three shells, and will tend to be copper-rich, and so on.

Such a variation in density of copper atoms with radial distance from the origin is analagous to the variation in density of the atoms about a given atom in a monatomic liquid. Debye and Menke[12] determined the radial distribution of the density of mercury atoms about a given atom in the liquid by x-ray methods. Their curve has been added in Fig. 6 with the dimensions adjusted so that the first peak coincides with the point representing the excess of copper atoms in the first shell. It is seen that the positive and negative portions of the curve coincide with the regions of copper defect and excess. The curve as drawn would represent the distribution in a liquid of "atoms" consisting of a copper atom plus half of its gold-rich shell of nearest neighbors. The distribution of copper atoms about a gold atom may thus be likened to a "melting" of the ordered structure to give a "liquid-like" arrangement of the atoms on the atomic sites.

This liquid-like dependence of the density of copper atoms on the radial distance is combined with the dependence on the lattice coordinates associated with ordering. The radial dependence becomes more important as the temperature is raised, and seems to predominate at temperatures as high as 550°C.

MEASUREMENT OF LONG-RANGE ORDER IN Cu_3Au

When an x-ray beam is reflected at the face of a single crystal by planes parallel to the face, the integrated intensity of the reflection is given by

$$E=K(N^2F^2/\mu)[(1+\cos^2 2\theta)/(2\sin 2\theta)]\exp(-2M),$$

where N = number of unit cells per unit volume, K = constant for given primary beam conditions, F = structure factor for the given reflection, μ = linear

TABLE I. Experimental values of short-range order parameters for Cu_3Au at temperatures above the critical temperature.

i	c_i	Coords.	Perfect order	405°C	460°C	550°C
1	12	1,1,0	−0.333	−0.152	−0.148	−0.131
2	6	2,0,0	1.00	0.186	0.172	0.105
3	24	2,1,1	−0.333	0.009	0.019	0.026
4	12	2,2,0	1.00	0.095	0.068	0.045
5	24	3,1,0	−0.333	−0.053	−0.049	−0.032
6	8	2,2,2	1.00	0.025	0.007	−0.009
7	48	3,2,1	−0.333	−0.016	−0.008	−0.003
8	6	4,0,0	1.00	0.048	0.042	0.019
9	12	3,3,0	−0.333	−0.026	−0.022	−0.011
	24	4,1,1	−0.333	0.011	0.020	0.007
10	24	4,2,0	1.00	0.026	0.025	0.007

[12] Debye and Menke, Physik. Zeits. 31, 797 (1930).

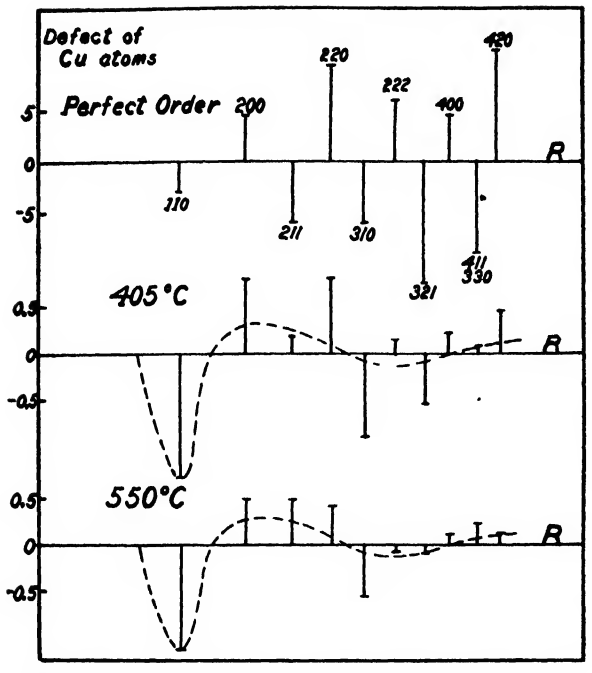


FIG. 6. The number of copper atoms in defect at various radial distances from a gold atom.

absorption coefficient of the crystal, $\exp(-2M)$ = correction factor for thermal vibration of atoms.

For a single crystal of Cu_3Au, with long-range order parameter S, the structure-factor for the super-lattice reflections is given by $F^2=(f_{Au}-f_{Cu})^2\cdot S^2$.

Wilchinsky[2] used the fundamental lines of Cu_3Au as intensity standards in his determination of S from powdered samples. Such a standardization is not possible when a single crystal is used, because the intensities of the strong fundamental reflections are decreased by an unpredictable amount by extinction effects. Hence an intensity standard was used; namely, the (600) reflection from a well-ground rocksalt cube face.

Measurements of the integrated intensities of super-lattice reflections from single crystals of Cu_3Au were made using the same apparatus as for the short-range order measurements outlined above. The entrance slit of the monochromator was narrowed to reduce the divergence of the primary beam to less than half a degree. Thus the volume element of reciprocal-lattice space shown in Fig. 2 was reduced to a thin plate-like element. This plate was swept through the region of a super-lattice reflection by a rotation of the single crystal specimen. The measured intensity was plotted against angle of rotation of the specimen. The area of the resultant peak was then proportional to the integrated intensity of the reflection. Similar measurements on the rocksalt reflection provided the standardization from which the value of S could be calculated. The results of two series of measurements are summarized in Fig. 7 and compared with the predictions of theory.[5]

The experimental results show agreement with the theory in the slow decrease in S and then the sharp fall to zero S at the critical temperature. The drop in S at the critical temperature was observed experimentally to take place within a temperature range of two or three degrees.

For such measurements, the single-crystal specimen was first heated at some temperature below the critical temperature for 70 to 100 hours. It was then found that, on raising or lowering the temperature, the integrated intensity reached a new equilibrium value in one hour or less. When the critical temperature was exceeded, the intensity of the super-lattice reflection became too small to measure within 15 minutes. Once the crystal was disordered, an annealing time of 70 hours or more was required to bring it to an equilibrium ordered state again.

These observations confirm the results and conclusions of Sykes and Evans,[13] who measured the electrical resistance of Cu_3Au as a function of time and temperature. They showed that two processes with very different time-constants could be distinguished; the slow growth of "anti-phase domains," and the much more rapid increase of order within the domains. Starting from a disordered sample, a long annealing time is necessary for the domains to grow large enough so that the presence of the domain boundaries has no appreciable effect on the electrical resistance or the integrated intensity of x-ray reflection. Once this domain size has been attained, an equilibrium state may be obtained rapidly at another temperature by increase or decrease of order within the domains.

A few observations have been made of the changes of integrated intensity, peak intensity, and half-width of super-lattice reflections from an initially disordered specimen heated below the critical temperature. A crystal, previously heated to 410°C to disorder it, was held at 360°C. After 15 minutes the integrated intensity of the reflection was about half of its maximum value, but the width of the reflection was much greater than

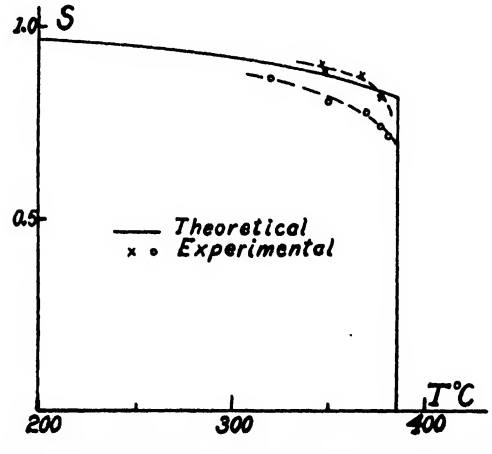


FIG. 7. Experimental values of the long-range order parameter, S, for Cu_3Au, compared with a theoretical curve.

[13] C. Sykes and H. Evans, J. Inst. Metals 58, 255 (1936).

for long annealing times. With further heating, the integrated intensity of the reflection increased more and more slowly. The rapid initial increase in intensity presumably corresponded to the propagation of order in the lattice outward from a number of nuclei until the domains so formed either coalesced or formed inter-domain boundaries. From a more detailed study of the shape and size of the super-lattice reflections at the end of this initial stage of growth, it should be possible to deduce the shape and size of the initially formed domains, and hence the directions of growth from the nuclei, and the number of nuclei.

The author wishes to express his indebtedness to Professor B. E. Warren, who suggested this research and provided valuable discussion and advice during the course of the work.

Integrating Crystal Detectors for High Energy Photons and Particles*,**

R. S. ALGER
Laboratory for Insulation Research, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received July 18, 1949)

Three of the alkali halides, LiF, KBr, and NaCl were examined as integrating detectors for high energy photons and particles. Dosage measurements were based on the number of color centers formed in single crystals under x- and cathode-ray bombardment. Exposures of 20 to 10^4r (30-kv x-rays) were measured photoelectrically and 10^5 to 10^7r (3 Mev x- and cathode rays) were measured optically. The crystals have advantages in measurements where wide range, high absorption, small size, and simplicity are desirable; however, their use is restricted by the length of the bleaching period required between exposures.

INTRODUCTION

CERTAIN of the effects produced in the alkali halides under high energy particle or photon bombardment, namely, color center formation and photoconductivity, suggest the possibility of a simple type of integrating detector which uses these materials. Quantitative measurements can be made, either of the optical absorption or the photoelectric currents, and the high absorption coefficients of the heavier crystals such as KBr would permit the detection of high energy particles with a relatively small volume of crystal. Consequently, three of the alkali halides, LiF, NaCl, and KBr were examined as to their characteristics as integrating detectors for x- and cathode rays up to energies of 3 Mev. Although, in this study, the emphasis was placed on experiments clarifying the usefulness of the crystals as detectors, data were obtained on the trapping and migration of electrons which, when further expanded, promise to give new information of general interest.

COLOR CENTERS IN THE ALKALI HALIDES

The following known facts concern the problem at hand. When the crystals are exposed to high energy photons or particles, they become colored by the *F*-absorption bands, which appear in the visible or ultraviolet region of the spectrum.[1] The *F*-centers responsible for this absorption seem to be electrons trapped at halogen ion vacancies;[2] such vacancies are always present in the alkali halides as Schottky defects.[3]

When x-ray photons transfer energy to the crystal, either by photoelectric absorption, Compton absorption, or pair production, primary electrons of high energy are released which dissipate their energy in secondary ionization processes, thus creating a number of conducting electrons and neutral halogen atoms. The electrons move through the crystal until they recombine with neutral halogen atoms or are trapped by negative ion vacancies to form *F*-centers. When crystals are bombarded with electrons, the final result is the same as with x-rays. Cathode and beta-rays may be stopped in the crystal and trapped as additional *F*-centers, but this contribution is negligible because of the space charge formed.

The *F*-center density can be calculated from the bell-shaped optical absorption band by an expression derived by Smakula.[4] Seitz[5] has simplified Smakula's equation, for the case where the half-band width is small compared to the frequency of the absorption maximum, to the form

$$N_0 f = 1.31\times 10^{17} K W \frac{n_0'}{(n_0'^2+2)^2} \qquad (1)$$

where N_0 = number of *F*-centers per cc, f = oscillator strength of the absorbing center, K = maximum optical absorption coefficient in cm^{-1}, n_0' = refractive index at K, W = interval from K to $K/2$ in e.v. If an oscillator strength of 0.7 is used for NaCl at room temperature, Eq. (1) reduces to

$$N_0 = CK. \qquad (2)$$

Values of the constant C listed by Hilsch and Pohl[6] for

* Sponsored by the ONR, the Army Signal Corps, and the Air Force under ONR Contract N5ori-07801.
** From a Master's Thesis in Electrical Engineering, M.I.T.
[1] R. W. Pohl, Proc. Phys. Soc. London **49**, E3 (1937).
[2] J. H. DeBoer, Rec. Trav. Chim. Pays-Bas **56**, 301 (1937).
[3] C. Wagner and W. Schottky, Zeits. f. physik. Chemie (B) **11**, 163 (1930).
[4] A. Smakula, Zeits. f. Physik **59**, 603 (1930).
[5] F. Seitz, *The Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940) p. 664.
[6] R. Hilsch and R. Pohl, Zeits. f. Physik **64**, 606 (1930).

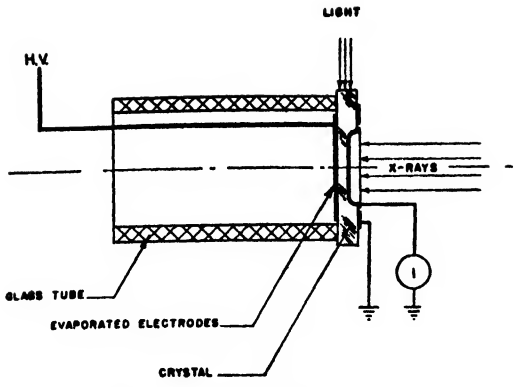


FIG. 1. Photoelectric dosimeter.

NaCl and KBr at room temperature are 3.6×10^{15} and 4.45×10^{15} respectively.

In addition to the principal absorption bands, for which the F-centers are responsible, a number of subsidiary bands have been observed at longer wavelengths, e.g., the F'-, R_1-, R_2-, and M-bands. Molnar[7] has measured the F-bands and the subsidiary bands for the majority of the common alkali halides.

When crystals colored with x-rays are heated or illuminated with light in the F-band, the whole F-band bleaches out uniformly; simultaneously some of the subsidiary bands may appear and a photo-current will be observed in case an external field is applied. Pohl and his co-workers,[1] who investigated the photoelectric processes extensively, found that one F-center was destroyed for every quantum of light absorbed. The conducting electrons move through the crystal until they are trapped by a vacancy or recombine with a halogen atom. The drift distance in the field direction depends on the strength of the electric field and the number of traps present. Pohl and his co-workers[1] designated the current due to photo-electrons as the "primary photoelectric current" whereas other currents resulting from the primary current flow were called "secondary photo-currents." These secondary effects occurred only above room temperature and were not observed in the experiments reported here. Since the electron drift distance is proportional to the field strength, the photo-current can be enhanced by the use of very strong electric fields. With photo-currents of the order of 10^{-10} amp., von Hippel[8] observed a photo-current saturation in thin KCl crystals beginning at a field strength of about 0.6 million volt/cm. Under more intense illumination thicker crystals break down electrically before a saturated photo-current can be obtained. At the saturation point the electron drift distance becomes equal to the thickness of the crystal, and all the electrons released in the crystal are swept to the anode.

[7] J. P. Molnar, thesis, Physics Department, M.I.T. (October, 1940).

[8] A. von, Hippel Phys. Rev. 54, 1096 (1938).

When a very weak source of illumination was used for short intervals of time, in which case the charge released was of the order of 10^{-12} coulomb, Pohl[1] found the primary photo-current remained nearly constant with time and had a magnitude proportional to the light intensity. When larger amounts of charge, e.g., 10^{-9} coulomb, are released, a space charge builds up in the crystal, which may nullify the external field, and no further photo-current is obtained. Because the range of the electrons is altered due to the field distortion, not all the photo-electrons contribute equally to the photo-current in this space charge region. Thus it becomes difficult to determine the F-center density from photo-electric measurements at high light intensities.

EXPERIMENTAL PROCEDURE

Crystal Preparation and Crystal Holders

For our photoelectric measurements NaCl and KBr crystals† were split to a size of $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{16}$ in. and hollowed on one side to keep the electric field low at the edge of the electrode. The cavities were produced with an automatic wet cloth lap. Gold electrodes were evaporated onto the polished crystals after they had been heated to 350°C in vacuum to expel surface moisture. Figure 1 gives a schematic view of a finished crystal mounted for photoelectric measurements. The glass tube, cemented to the crystal with polybutyl methacrylate, supports the sample and prevents surface flashover. X-rays are admitted through the concave electrode and the light is focused on the edge of the crystal.

LiF, NaCl, and KBr samples, $1\times1\times0.1$ in. and $1\times1\times0.01$ in., were prepared with optical surfaces for the F-band absorption measurements.

Exposure Methods

Two radiation sources were used; a Machlett x-ray diffraction tube at 30 to 45 kv with a copper target and a 1 mil thick nickel filter, and a 3-million volt source for x- and cathode rays (van de Graaff generator).

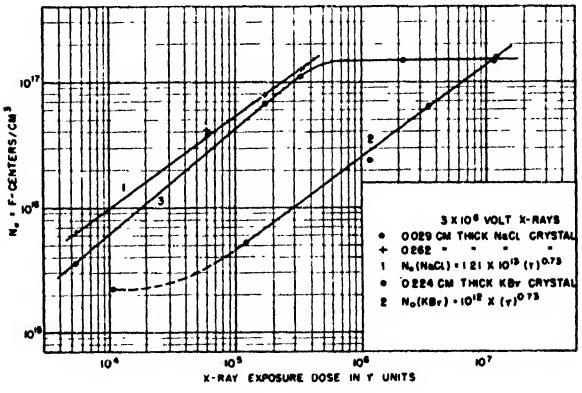


FIG. 2. F-center concentration as function of x-ray exposure.

† Harshaw Chemical Company, Cleveland, Ohio.

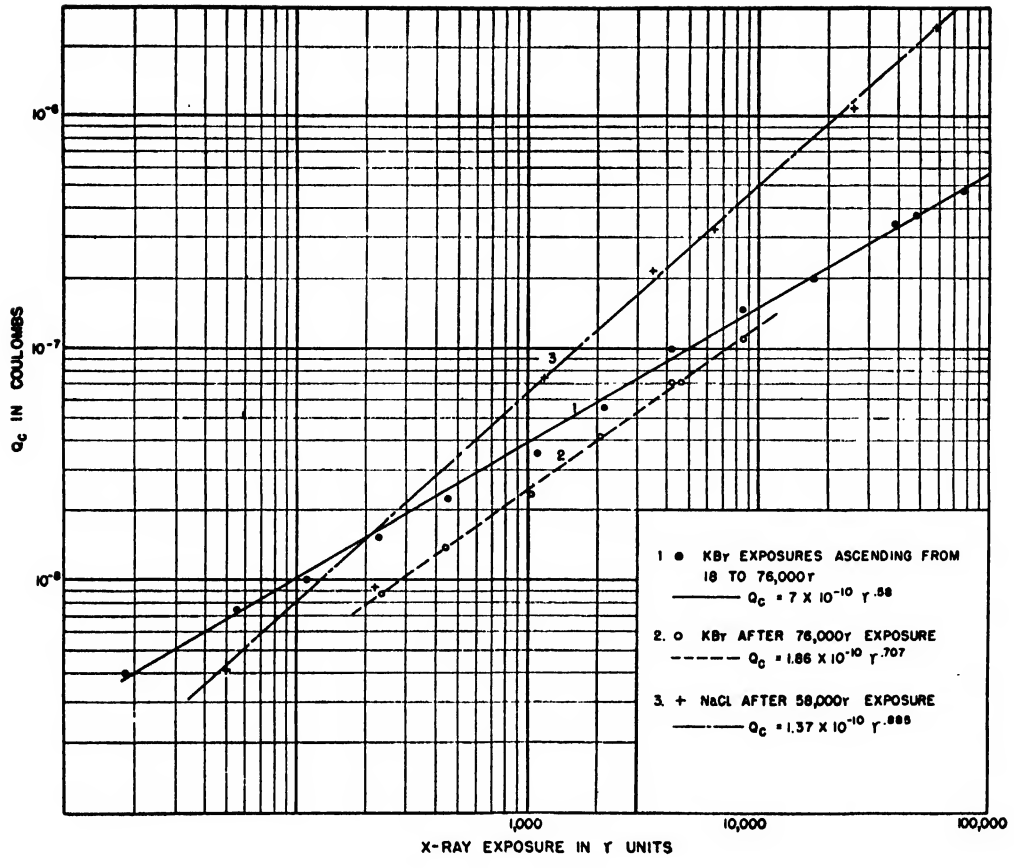


FIG. 3. Charge bleached from crystals as function of x-ray exposure.

All x-ray exposures were measured in roentgens (r units) with air ionization chambers. The chamber used with the diffraction tube was copied, with some change of dimensions, from a design of Taylor and Stoneburner.[9] Exposures to 3-million volt x-rays were measured with a roentgen meter built into the van de Graaff generator. While the relative values measured with this meter are accurate to better than 5 percent, the actual calibration in roentgens may be in error by a factor of 2. Cathode-ray exposures were measured in joules and converted to roentgen-equivalent-physical (rep) units in air for comparison with the x-ray doses. (1r$\simeq$83.8 ergs/g of air at standard conditions and 1 rep=83.0 ergs/g of material.) Due to the simplifying assumption of a uniform ionization intensity and the unknown distribution of cathode rays over the target plate, the accuracy of the rep values is probably less than that of the r.

Since the production of F-centers depends on ionization, the F-center concentration will vary throughout the crystal according to the ionization intensity. For the 30 to 45-kv x-rays the ionization is most intense at the incident surface and decreases rapidly in accordance with the large coefficient of absorption for the crystals. At 30 kv and exposures up to 10^4 r, the visible coloring extended less than 1 mm into the KBr crystals. Photoelectrons resulting from the absorption of 3-Mev photons have an appreciable velocity in the direction of the x-ray beam; therefore, the peak ionization intensity occurs beyond the incident surface and may well be at the exit surface of a thin crystal. Consequently a given x-ray exposure below the saturation value might produce a smaller average F-center concentration in a thin crystal than in a slightly thicker one because the latter contains more of the peak ionization.

Photoelectric and Absorption Measurements

The coloration produced in the crystals by exposure to an x-ray dose of from 20 to 10^4 roentgens was measured photoelectrically by the charge Q_c traversing the crystal under strong illumination. After the crystal had been colored (Fig. 1), an electric field of about 10^5 volts/cm was applied, the crystal illuminated and the photo-current recorded as function of time. The space charge formed during the measurement of Q_c cuts off the photo-current before all of the F-centers are removed; the remaining centers can be bleached by illuminating the crystal until no photo-current is obtained with either polarity of the external field.

The coloration of crystals given larger exposures (10^4 to 10^7 r or rep, of 3-million volt rays) can be measured optically. Values of the maximum absorption co-

[9] L. S. Taylor and C. F. Stoneburner, Bur. Stand. J. Research 9, 769 (1932).

efficient K, from which N_0 is obtained by Eq. (2), were measured with a Model 12M Cary recording spectrophotometer. The minimum measurable transmission was about 0.03 percent, and no F-center bleaching due to the scanning beam could be detected. Absorption measurements covered the region from 3400 to 7200A in KBr and NaCl, and from 2100 to 3400A in LiF. Desiccated holders protected optical surfaces of the hygroscopic crystals from atmospheric moisture.

The KBr and NaCl crystals were bleached with white light between experiments, while a 4-watt germicidal lamp, which has its major intensity concentrated in the 2537-Hg line, was used to bleach LiF. A residual absorption was bleached out by heat.

Parameters influencing both N_0 and Q_c are: besides the crystal material and prehistory, the energy and intensity of the radiation, the temperature and time of storage, the voltage and crystal thickness used. The effects of each of these variables were examined in turn.

RESULTS

Effect of Exposure

The relation between the F-center concentration N_0 in KBr and NaCl and the x-ray exposure in roentgens over the range from 10^4 to 10^7 r is shown in Fig. 2. N_0 increases with exposure until an equilibrium condition is reached (curve 3), where a saturation value of 1.5×10^{17} F-centers per cc was obtained for NaCl. This saturation can be measured only in thin crystals; in thicker crystals (curves 1 and 2) practically all the light is already absorbed at lower concentrations, and the sensitivity range of the spectrophotometer is exceeded. Over the straight line portion of the log-log curves 2 and 3 the F-center concentration fits the empirical relation

$$N_0 = C_1 r^{0.73}, \tag{3}$$

where $C_1 = 1.0 \times 10^{12}$ for KBr and 1.2×10^{13} for NaCl. Since the F-band was barely detectable in KBr at $r = 10^4$, this point may be considerably in error.

Typical photoelectric measurements of Q_c for 30-kv x-rays and exposures ranging from 20 to 10^4 r are given for KBr and NaCl in Fig. 3. The empirical equations obtained between Q_c and the x-ray exposure in roentgens are similar to Eq. (3).

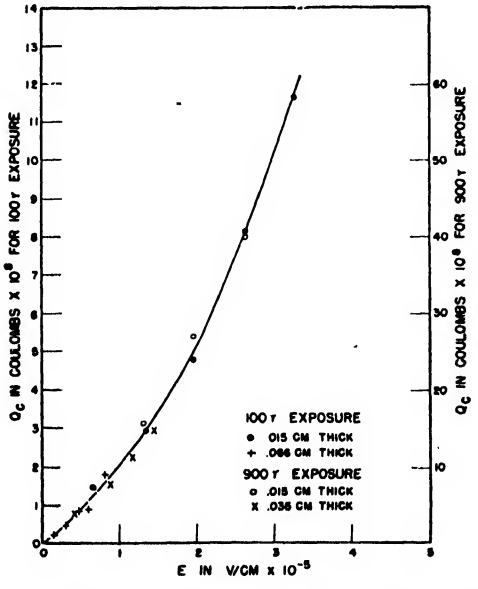


FIG. 4. Effect of applied field on charge bleached from KBr crystals.

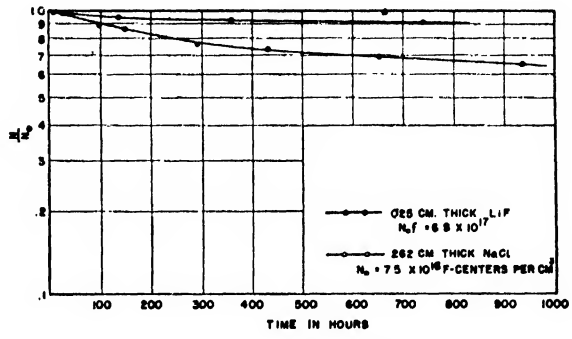


FIG. 5. F-center decay at 21°C due to thermal release.

For KBr (curve 1) $Q_c = 7 \times 10^{-10}$ $r^{0.68}$, KBr (curve 2) $Q_c = 1.86 \times 10^{-10}$ $r^{0.707}$, NaCl (curve 3) $Q_c = 1.37 \times 10^{-10}$ $r^{0.885}$.

It becomes apparent from curves 1 and 2 that a preceding exposure can affect the characteristics obtained in later measurements (see later section on Polarization).

Influence of Crystal Type and Thickness

The crystal material has a pronounced effect on the F-center density N_0 and charge Q_c obtained for a given x-ray exposure (Figs. 2 and 3). When the crystals are exposed to x-rays, the F-center formation rate (i.e., N_0/roentgen) diminishes in the order LiF→NaCl→KBr. Comparing LiF with NaCl for an exposure of 10^7 rep, the ratio $N_0 f_{(\mathrm{LiF})}/N_0 f_{(\mathrm{NaCl})} = 4.6$ was found (see Eq. (1)). A 1 mil thick crystal of LiF became practically opaque under a 3-million volt x-ray bombardment before the F-band of a similarly treated KBr crystal was readily detectable. In Fig. 2 the F-center density $N_{0(\mathrm{NaCl})}$ is seen to be greater than $10 N_{0(\mathrm{KBr})}$ for identical x-ray exposures below the saturation value.

Since the optical measurements show more F-centers formed in NaCl than KBr, it would be expected that for crystals of the same thickness and subjected to the same electric field, a larger charge Q_c would be obtained for NaCl. A comparison of the curves in Fig. 3 shows that this is the case; $Q_{c(\mathrm{NaCl})}/Q_{c(\mathrm{KBr})}$ varied from about 2 to 4. Selection of the proper crystal material provides some control over sensitivity and usable range of these integrating detectors.

N_0 is slightly higher in a thicker crystal for a given exposure of 3-Mev x-rays (Fig. 2), which agrees with

the result predicted by the distribution of the ionization intensity. This variation should affect N_0 only before the crystal becomes saturated. Figure 4 shows Q_c as a function of the applied field strength for 3-crystal thicknesses. No definite thickness effect is apparent over the limited region covered.

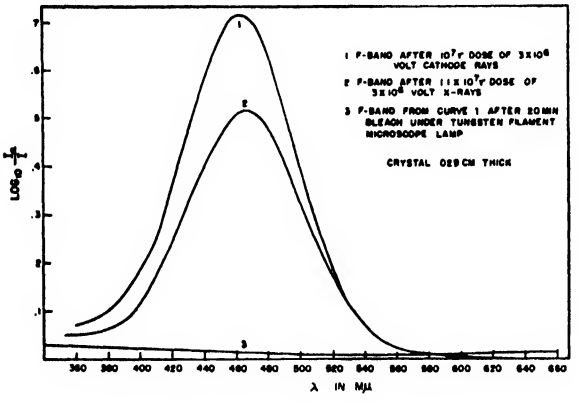


FIG. 6. Bleaching of F-centers in NaCl.

Thermal Bleaching

In alkali halide crystals colored by x-rays, some thermal bleaching occurs at room temperature. The bleaching rate may be very slow, e.g., Joffé observed some color centers remaining in an x-ray colored NaCl crystal after it had been stored in the dark for about 20 years.[10] The heavier crystals, with their lower activation energy, bleach more rapidly than the crystals composed of the lighter elements. Since the thermal bleaching would produce an error in dosimeter readings, a series of measurements were made on colored NaCl, LiF, and KBr crystals to determine the bleaching rate. The results for NaCl colored by 3-million volt x-rays, and LiF colored by 3-million volt cathode rays are given in Fig. 5 which shows N/N_0 plotted as a function of time. (N = the F-center concentration at any time t and N_0 = the initial concentration.) The decay curves are not simply exponential, but a half-life of about 0.87 year fits the LiF characteristic quite well.

A consideration of the mechanism of decay indicates that a more complicated expression would be required to describe completely the decrease in the F-center density as a function of time. When the electron is thermally released from an F-center, it has the possibility of being captured by a neutral halogen atom or another halogen lattice vacancy. Each time an electron is captured by a neutral halogen atom the number of neutral halogen traps is decreased by one, and at the same time the ionized F-center increases the number of lattice vacancies by one. Consequently, as bleaching continues, the ratio of halogen vacancies to halogen neutral atoms increases, and the probability that the thermally released electron will be captured by a lattice vacancy to reform an F-center also increases. Therefore the F-center density decreases more rapidly the higher the coloration. Thus the thermal bleaching produces a larger error in the optical measurements where high F-center densities are employed than in the photoelectric measurements. In the case of photoelectric measurements with KBr, the time between x-ray exposure and the photoelectric measurement was allowed to vary from a few seconds up to 20 hr. at room temperature, and no decrease in the charge Q_c was observed. However, a slight decrease in the optical density of the F-band in highly colored KBr and NaCl can be detected after several hours.

Optical Bleaching

If the crystals are to be used more than once, the remaining color centers must be removed and the space charge redistributed between measurements. The required bleaching period depends on the F-center density, bleaching light intensity, and the crystal material. For example, 5 minutes of illumination were sufficient to bleach a KBr crystal exposed to a 100-r dose, while an NaCl crystal given the same dose required 30 minutes of illumination.

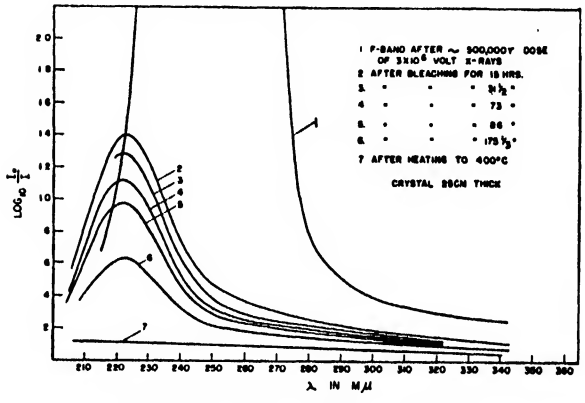


FIG. 7. Bleaching of F-band in LiF.

The optical measurements show the results of photo- and thermal bleaching more directly. Curve 3 of Fig. 6 gives the result of 30-minutes' bleaching with white light of the F-band absorption (curve 1) in an NaCl crystal exposed to approximately 10^7 rep. The F-band has been completely bleached, and no R- or M-bands are present; however, the absorption through the entire spectrum has increased from 0.015 to 0.02 density units.

The bleaching of LiF produced an effect not encountered thus far in the other crystals. An LiF crystal with an F-band absorption well above the upper limit of measurement (Fig. 7, curve 1) was exposed to the light from a 4-watt germicidal lamp for 175 hours. After the first 15 hr., the F-band was almost completely bleached except for a small peak centered at 2230A (curve 2). With continued bleaching the subsidiary

[10] A. F. Joffé, *The Physics of Crystals* (McGraw-Hill Book Company, Inc., New York, 1928) p. 128.

peak decayed symmetrically as shown in curves 2 to 6. Normally, illumination with any frequency in the F-band causes the whole band to decay symmetrically with the possibility of the F'-, R-, and M-bands appearing at longer wave-lengths. The auxiliary band at 2230A is singularly different from the usually subsidiary bands since it is on the short wave-length side of the F-band. When the crystal was heated to 400°C and allowed to cool, the F-band and the subsidiary peak were eliminated (curve 7). Once more, the final curve (7) exhibits some neutral absorption over the entire spectrum investigated; it may be due to surface deterioration.

Effect of Vacancies on Performance

Since the number of vacancies is dependent on the crystal temperature, the vacancy concentration can be increased by heating the crystals almost to the melting point and cooling them rapidly to room temperature. The number of vacancies frozen into the crystals quenched in this way will correspond to some intermediate temperature. Many of the vacancies appear to be paired up, i.e., halogen and alkali vacancies are attached to each other. A method for determining the number of paired vacancies from low frequency dispersion measurements has been described by Breckenridge.[11]

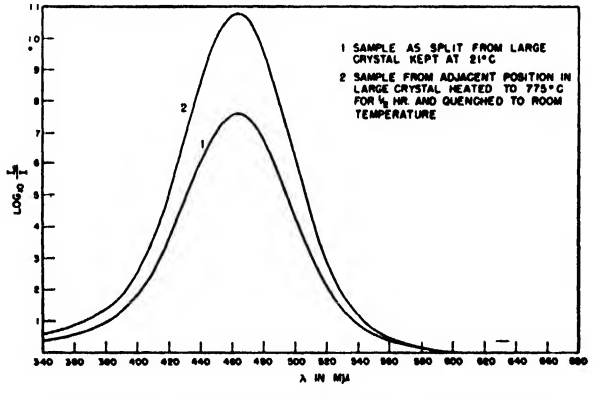


FIG. 8. Effect of vacancies on F-band in NaCl crystals.

If the dielectric loss ($\tan\delta$) is measured at a constant frequency of about 10^4 cycles/sec. over a range of temperatures, the alkali halide crystals containing sufficient vacancies exhibit a peak in the $\tan\delta$-curve at a temperature characteristic of the crystal. The number of paired vacancies can be determined from the height of the $\tan\delta$-peak by the following equation:[11]

$$n=\tan\delta_{\max}\frac{(18\epsilon_0 k)T\kappa_\infty'}{(Zea)^2(\kappa_\infty'+2)^2}, \qquad (4)$$

where $\epsilon_0=8.856\times10^{-12}$ farad/meter, $e=1.59\times10^{-19}$ coulomb, $k=1.38\times10^{-23}$ joule/degree centigrade, $\kappa_\infty'\simeq\kappa'=5.9$ for NaCl. By measuring $\tan\delta$ as a function of temperature before and after a crystal has been heat-treated, the number of paired vacancies produced by the heat treatment can be determined from the above relation. Crystals of LiF, KBr, and NaCl, which had been heated to increase the number of vacancies present, were compared with unheated crystals by measuring both the number of paired vacancies and F-centers formed in the crystals with a given dose of x-rays. Since electrodes were required for the dispersion measurements, two sets of crystals were used, one for the electrical and one for the optical measurements. The crystals to be compared were split from adjacent positions in the original crystal. NaCl was kept at 775°C for 30 minutes, KBr at 710°C for 30 minutes, and LiF at 825°C for 60 minutes and all were quenched to room temperature. This heat treatment produced only a slight hump in the $\tan\delta$-curves for KBr and LiF, and a more pronounced jump in NaCl, which, according to Eq. (4) corresponded to $n=1.5\times10^{15}$ vacancy pairs/cm.[3]

A 250,000-r dose of 3-million volt x-rays produced a stronger coloration in the heated samples; the effect was larger in NaCl than in either LiF or KBr as was to be expected from the $\tan\delta$-characteristics. Figure 8 shows the result for NaCl. The values found for the F-center densities per cc were:

$$N_{01}=5.8\times10^{16} \text{ for the heated samples}$$
$$\frac{N_{02}=4.1\times10^{16}}{\Delta N_0=1.7\times10^{16}} \text{ for the unheated samples.}$$

The value of N_{01} is a factor of 40 higher than the number n of paired holes obtained from the dispersion measurements (see Eq. (4)), although the F-center concentration is less than half the saturation concentration shown in Fig. 5. If, as Seitz suggests,[12] the vacancies are completely paired after the crystal stands for a few days at room temperature, the maximum number of F-centers should not exceed the number of

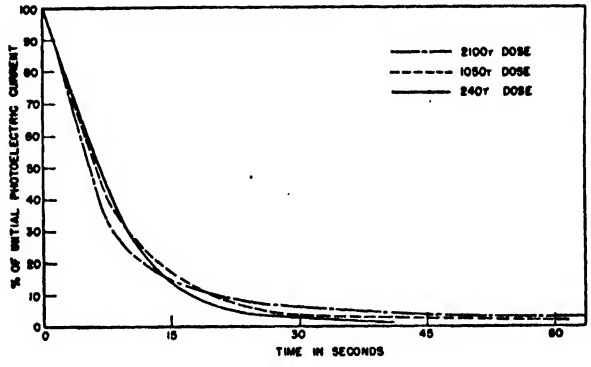


FIG. 9. Reduction in photo-current in KBr due to polarization.

[11] R. G. Breckenridge, Technical Report XI, Lab. Ins. Res., M.I.T. (April, 1948).

[12] F. Seitz, Rev. Mod. Phys. 18, 384 (1946).

paired vacancies if no new vacancies are created by the x-ray bombardment. The ratio of $n/N_0=0.025$ obtained in this experiment would indicate that only a small percent of the vacancies were paired.

The rate of F-center formation was increased in the crystals containing the larger concentration of vacancies. Consequently, crystals containing the largest possible number of vacancies might make the most sensitive detectors, but the effect is not very large.

Polarization

In photoelectric measurements, the space charge formed has a pronounced effect on the value of Q_e observed. When the crystal is illuminated in a strong electric field the electrons released from the F-centers will move toward the anode. Some electrons will be removed from the crystal while others are retrapped to form new F-centers or trapped by neutral halogen atoms to form ions. A positive space charge, presumably of neutral halogen atoms, develops in the vicinity of the cathode, and reduces the primary photo-current essentially to zero as the applied field is nullified. This behavior is illustrated by the curves of photo-current *vs.* illumination time in Fig. 9. Some of the reduction in primary photo-current is due to the loss of charge carriers to halogen neutral traps; however, the reduction in current shown in Fig. 9 is primarily related to space charge.

Increasing the F-center concentration results in a greater space charge density and consequently a steeper potential gradient at the cathode. When the gradient becomes sufficiently steep, electrons can enter the crystal and a semicontinuous photo-current is observed. After the electrons have entered the crystal they are indistinguishable from the electrons released from F-centers. Gradually the competing neutral halogen traps reduce the F-center concentration thereby preventing continuous photo-conductivity. In KBr crystals exposed to 10^7 rep of 3 million-volt cathode rays the total photo-current may drop by a factor of 3 during half an hour with the rate diminishing with time as the ratio of halogen vacancies to halogen neutral atoms increases. At high F-center concentrations Q_e is dominated by the rate of trapping by halogen neutral atoms, while at low F-center concentrations, the polarization predominates. With the equipment used in these experiments a practical upper limit of about 50,000 r for KBr was dictated by the bleaching time required to reduce the total current to zero. Bleaching the crystal with light does not restore the crystal completely to the initial condition of electrical neutrality. After the initial exposure r_1, with subsequent bleach and polarization, a series of reproducible values of Q_e can be obtained for exposures less than r_1. However, with $r_n > r_1$, a change occurs in the charge distribution when Q_{e_n} is measured and a pronounced decrease in Q_e is observed for all future exposures less than r_1. This effect is illustrated in Fig. 3. Curve 1 was obtained for exposures progressing from 18 to 78,000 r. After the 78,000-r exposure, smaller exposures fall along curve 2. The same effect was observed with NaCl before curve 3 was obtained. The polarization effects are being studied in more detail.

CONCLUSIONS

The results obtained demonstrate that the color centers in the alkali halides can be used to measure radiation or particle exposures. However, there are a number of restrictions which limit the usefulness of crystal dosimeters.

The main advantages of the crystal dosimeters are: small size, high absorption, wide range, and simplicity of the detecting element. Although the smallest crystals used in the experiments were $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{16}$ in. in size, smaller crystals could be used if desired. When compared to gas-filled detectors such as ionization chambers, the higher absorption coefficient of the crystals is an advantage, especially when high energy radiations are being detected. Exposures of 20 to 10^7 r measured with KBr crystals illustrate the wide range possible. The crystal material offers some choice in the range and sensitivity that can be obtained. Based on the optical measurements, the crystals listed in descending order of sensitivity are: LiF, NaCl, and KBr.

The bleaching period required to eliminate the F-centers between successive measurements is a major limitation of the speed with which measurements can be repeated.

The polarization effects complicate the photoelectric measurements and make frequent calibrations necessary; consequently the optical method is more attractive for exposures above 10^5 r.

The vacancy concentration has only a slight effect on the F-center formation rate, and thermal bleaching at room temperature can be neglected except in the case of large F-center concentrations or long storage periods.

The author is indebted to Professor A. von Hippel under whose supervision this work was carried out, to Professor G. J. Trump for the use of the 3-million volt generator, to K. A. Wright who operated the generator and to M. A. Gilleo for assistance with the optical measurements.

Thermionic Emission from Oxide Cathodes: Retarding and Accelerating Fields*,†

C. S. HUNG‡
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received August 3, 1949)

An experimental investigation of the thermionic emission from oxide-coated cathodes is presented in which the applied voltage range extends from the retarding potential region through zero field and up to an average accelerating field of 50,000 volts per cm. The retarding potential range yields information concerning the energy distribution of the electrons emitted, and indicates an excellent agreement between theory and experiment for high retarding fields. Near zero field, cathode inhomogeneity or patch effect may be responsible for the deviations found. Consideration is given to the influence of a possible reflection effect, but theoretical calculation indicates that with the large area cathode used, the simultaneous reflection at the emitter and the collector surfaces so alter the observed current voltage relationship, that the existence or non-existence of a reflection effect cannot be determined by the experiment. Results shown in the accelerating field region up to 50,000 volts per cm are accounted for on the bases of two assumptions which are (1) patch effect for small accelerating fields and (2), the intensification of the field at sharp points expected in view of the inherent roughness of the cathodes studied. The deviation of the experimental results from those predicted by the Schottky mirror image theory is explained over the whole range of observation by these two assumptions.

INTRODUCTION

THE object of this work is to study two of the properties of the surface potential barrier of an oxide cathode. The first is the reflection effect on electrons crossing this barrier, and the second is the lowering of this barrier height in the presence of an external electric field. The surface potential barrier of a uniform thermionic emitter is usually assumed, for distances large compared to the lattice constant, to be due to the mirror image force alone. The quantum-mechanical reflection coefficient for an image force barrier, according to Frank and Young,[1] is only six percent for the low energy electrons and is a slowly varying function of the electron energy in the direction normal to the cathode surface. One way of determining the reflection coefficient experimentally is to measure the electron current to the anode as a function of the retarding potential of the anode relative to the cathode. The thermal equilibrium energy distribution of electrons in the emitter, with energies above the height of this barrier, is known to be Maxwellian since the height of the barrier from the μ-level is always large compared to kT at all operating temperatures of the cathode. This retarding field current, for cylindrical electrodes and assuming zero reflection coefficient is, as given by Schottky:[2]

$$J_v = J_0[2/(\pi)^{\frac{1}{2}}]\left\{e^{-n}\int_0^{(\lambda n)^{\frac{1}{2}}} \exp(-\theta^2 x^2)dx + \int_{(\lambda n)^{\frac{1}{2}}}^{\infty} \exp(-x^2)dx\right\}, \quad (1)$$

where J_v, and J_0 denote the currents per unit length of the cathode when the retarding potentials are V and zero respectively. And $n=eV/kT$, $\theta=r/R$, $\lambda=R^2/(R^2-r^2)$, where r, R denote the radii of the cathode and the anode respectively. The J_v/J_0 *vs.* V/T curves are independent of T under the conditions of this equation. Using this retarding potential method, Nottingham[3] found that for pure tungsten and thoriated tungsten of all degrees of coverage at the same temperature, the plots of $\log(i_v/i_0)$ against V/T all coincide, but the plots for different temperatures differ from each other and from the theoretical curve assuming no reflection

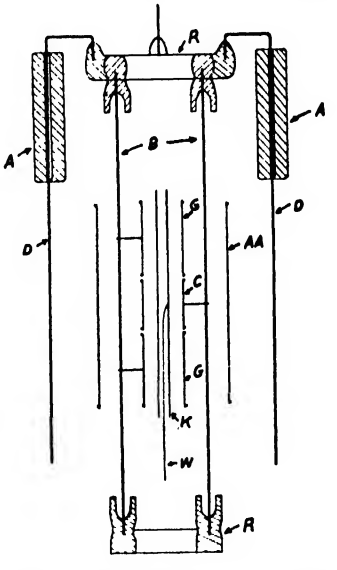


FIG. 1. The sliding anode structure. Fixed elements: K—cathode, 0.2 cm in diameter and 4.5 cm long; AA—auxiliary anode, 2.5 cm in diameter; W—0.002-in. tungsten wire as one element of the thermocouple; A—glass sleeves for guiding the motion of the sliding structure. Sliding structure includes: G—guard rings, 0.8 cm in diameter; C—collector, 0.8 cm in diameter; R—glass rings as insulation and support; B—6 tungsten rods sealed in R for supporting the guard rings and the collector; D—tungsten rods sealed in R, sliding in A. The sliding structure is in the normal position for emission measurements.

* From a thesis paper submitted to the Department of Physics, M.I.T., in partial fulfillment of the requirements for the degree of Doctor of Science.

† This work was done with the support extended to the M.I.T., Research Laboratory of Electronics, jointly by the Army Signal Corps, the Navy Department (ONR), and the Air Force (Air Material Command).

‡ Now at the Department of Physics, Purdue University.

[1] N. H. Frank and L. A. Young, Phys. Rev. 38, 80 (1931).

[2] W. Schottky, Ann. d. Physik 44, 1011 (1914).

[3] W. B. Nottingham, Phys. Rev. 49, 78 (1936).

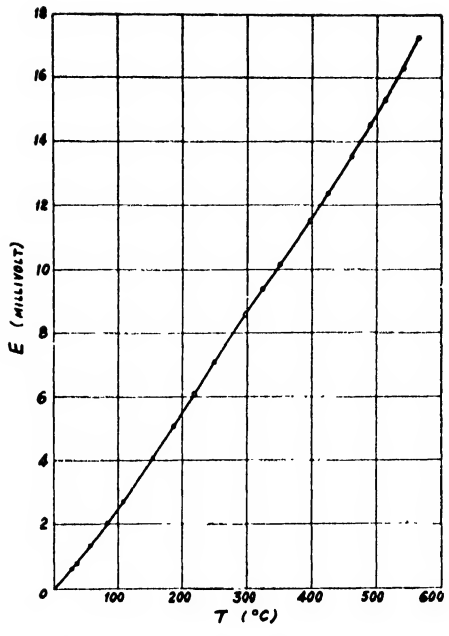


FIG. 2. Calibration curve of tungsten-nickel thermocouple. T is temperature of hot junction above that of cold junction (at 25°C).

as given by Eq. (1). A reflection coefficient of the form

$$R(p_z)=\exp(-p_z^2/2m\omega) \tag{2}$$

was proposed, where p_z denotes the momentum normal to the cathode surface that the electrons would have with zero applied field at distances beyond the range of mirror image forces. With the constant $k/\omega=4.5\times10^{-4}$ (deg.)$^{-1}$, all his data could be explained. This result was interpreted to mean that this reflection coefficient is probably related to the image field and not the surface patches, and therefore a reflection effect of at least this amount should exist at the surfaces of all electrodes including the anodes and the oxide cathode. Experiments on oxide cathodes by Heinze and Hass,[4] Fan,[5] and Brown,[6] have shown agreement with Eq. (1) for high retarding potentials. Deviations at low retarding potentials were reported and demand further study.

The lowering of the image force barrier in an external field is, according to Schottky,[2]

$$\Delta\phi=(eE)^{\frac{1}{2}} \tag{3}$$

and the anode current in an accelerating field is

$$J_v=J_0\exp[(eE)^{\frac{1}{2}}\cdot e/kT] \tag{4}$$

where E is the accelerating field at the cathode surface. Equation (4) is exact in the case of parallel plane electrodes and is also valid in the case of cylindrical electrodes except at very low anode voltages. Investigations on tungsten and thoriated tungsten cathodes by Nottingham[3] have shown agreement with Eq. (4) except for deviations at low accelerating potentials which have been explained satisfactorily by the patch theory. Measurements on oxide cathodes by Fan,[5] Mahlman,[7] Sproull,[8] and others have shown that the slopes of the Schottky plots, $\log(J_v/J_0)$ *vs.* $(E)^{\frac{1}{2}}/T$, in accelerating fields up to 10,000 volt/cm are 2 to 6 times higher than

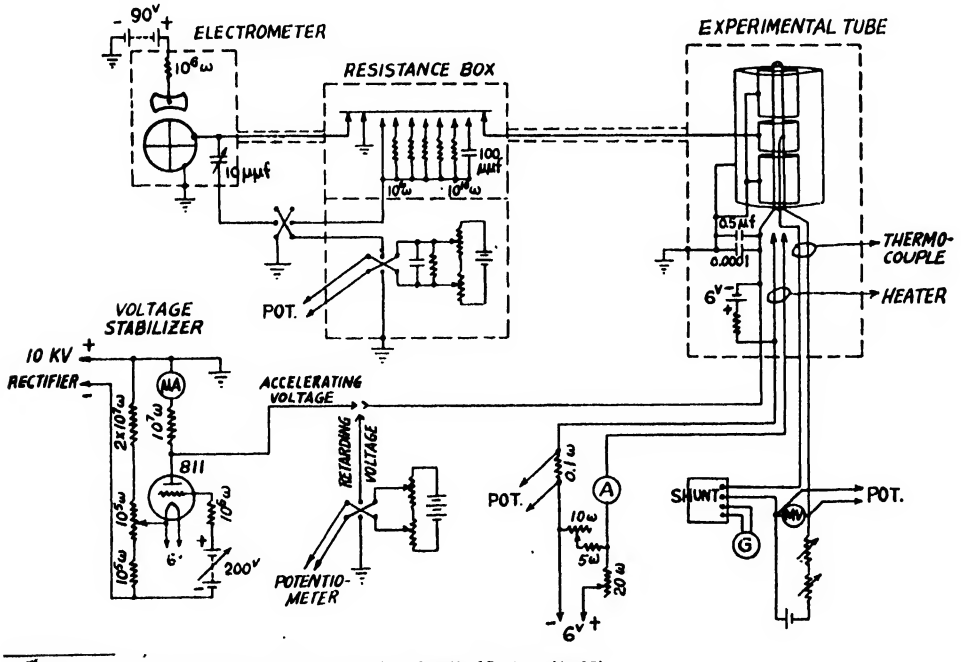


FIG. 3. The measuring circuit.

[4] W. Heinze and W. Hass, Zeits. f. tech. Physik **19**, 166 (1938).
[5] H. Y. Fan, J. App. Phys. **14**, 552 (1943).
[6] B. B. Brown, M. I. T. thesis, (1942).
[7] G. W. Mahlman, M. I. T. thesis (1948).
[8] R. L. Sproull, Phys. Rev. **67**, 166 (1945).

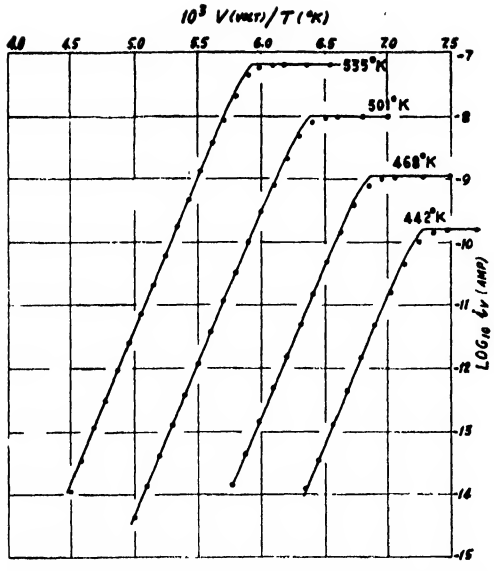


FIG. 4. Retarding field plots. Circles indicate the experimental points, and solid line the theoretical Schottky curves.

the theoretical image theory line. This effect was usually attributed to the patches, and it was suggested that the theoretical line might be approached at high enough accelerating fields.

In making further studies of these properties of the surface potential barrier, we notice that there are two more defects of the oxide cathode which should be considered: namely, the non-uniformity of the work function of the anode due to contamination, and the roughness of the oxide cathode surface. In this work we have managed to keep the anode free from the materials evaporated from the cathode so that the anode work function was uniform and did not change with the anode current. The roughness of the oxide cathode surface has been taken account of in analyzing our results.

EXPERIMENTAL SET-UP

The Experimental Tube

The main feature in the design of the experimental tube involved the use of a sliding anode structure in order to keep the anodes clean. A drawing of the sliding structure is shown in Fig. 1. The cathode (K) and the auxiliarly anode (AA) are fixed in position while the main anodes, the collector (C) and the guard rings (G), are removable. During the processing of the tube, the main anodes (GCG) were drawn up and were kept hot during the breakdown and the activation of the cathode so that materials evaporated from the cathode did not contaminate them. The auxiliary anode served as anode for drawing emission during activation. After the tube was sealed off, when the main anodes were suspected of being dirty, they were drawn up and heated to high temperatures to drive off the deposits. The collector and the guard rings were supported by a "bird-cage-like" structure, which is restricted to up and down motion through the use of the glass sleeves (A). The main anodes and the auxiliary anode were made of tantalum which does not show "anode effect"[3] after it has been properly cleaned.

The cathode sleeve was made of 2-mm diameter grade A nickel tubing. The cathode temperature was determined by the thermocouple method as used by Fan[5] and others. A 2-mil tungsten wire was welded to the center of the inside of the cathode sleeve. The point of welding served as the hot junction. The tungsten wire and a 5-mil nickel wire welded to the end of the cathode sleeve were brought out through a separate press. This press was immersed in an oil bath to keep the cold junction temperature constant. The disturbance of the cathode sleeve temperature[6] by this tungsten wire may be neglected. The temperature of the coating surface is believed to be within a few degrees of the sleeve temperature at all operating temperatures.[9] The tungsten-nickel thermocouple was calibrated against Chromel-P-Alumel thermocouple and checked against the readings of a glass thermometer for every hundred degrees. The calibration curve is shown in Fig. 2.

The main anodes and the auxiliary anode were chemically cleaned and induction heated to 1600°C for a prolonged period.

The cathode coating consisted of $(BaSr)CO_3$ sprayed onto the sleeve to an amount of 5 mg/cm^2. The tube was baked out thoroughly at 500°C, and after this the cathode was broken down and activated. The pressure of the tube after seal-off was around 10^{-8} mm Hg with the use of getter.

The Measuring Circuit

The measuring circuit is as shown in Fig. 3. The experimental tube was shielded in a copper box against disturbances. All leads coming out of the box were shielded. The heater current was measured by the

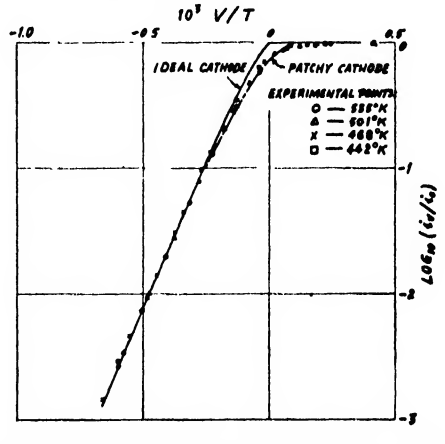


FIG. 5. Retarding field plots matched together. Solid line—theoretical Schottky curve; dot-dash line—calculated curve for a patch model at 500°K assuming no reflection.

[9] G. E. Moore and H. W. Allisón, J. App. Phys. 12, 431 (1941).

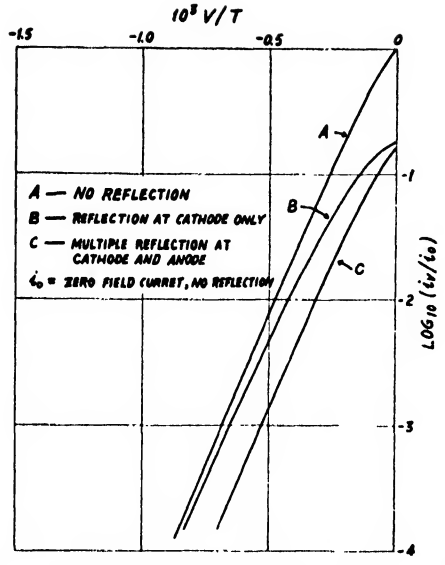


FIG. 6. Calculated retarding field plots for a uniform cathode with different assumptions of reflection effect.

voltage drop across a 0.1 ohm resistance with a Leeds and Northrup K-2 potentiometer. The negative lead of the heater was made 6 volts positive with respect to the cathode to suppress electron emission from the heater. The connections for thermocouple e.m.f. measurement were all copper wires from the cold junction to the meter. A change of cathode temperature by 0.2°C, resulting in a change of thermocouple e.m.f. of 0.007 mv, was indicated by a deflection of 0.2 mm on the galvanometer scale. The cathode was grounded through a 0.5 μf condenser in parallel with a 100 $\mu\mu$f mica condenser to shunt off electrical disturbances. The anodes were grounded and the cathode voltage was variable. Retarding voltages were obtained from dry cells and measured with the K-2 potentiometer. Accelerating voltages below 200 volts were obtained from dry cells, while those from 200 up to 8000 volts were obtained from a high voltage rectifier and controlled by a voltage stabilizer. The voltage stabilizer employed an amplification factor bridge circuit with an 811 tube. When properly adjusted, the variation in output voltage of the stabilizer was below 1 percent of the input variation for all load currents up to 1 μamp. The guard rings and the auxiliary anode were connected directly to ground, while the collector lead was led to the current measuring equipment through a coaxial copper tubing which was grounded. The wire was insulated from the tubing using polystyrene disks. The current measuring equipment consisted of a resistance and balancing e.m.f. circuit and a Compton quadrant electrometer. The voltage drop across a high resistance or a condenser due to the collector current was balanced out, as indicated by the zero deflection of the electrometer, and measured with the potentiometer. Vacuum tube platinum resistors up to 10^{10} ohms were used for current measurement down to 10^{-12} amp. and the charge accumulation method employing a 100 $\mu\mu$f air condenser, and a 10 $\mu\mu$f air condenser (built in the electrometer) were used for current measurement down to 10^{-14} amp.

RETARDING FIELD STUDIES

Retarding Field Current from a Non-Uniform Cathode

It is known that because of the different crystal faces, the crystal defects, and the various amounts of foreign matter adsorbed on the crystal faces, the work function of an oxide cathode is not uniform but varies over short distances along its surface. This non-uniformity of the cathode work function, the patches, causes the retarding field current to be different from

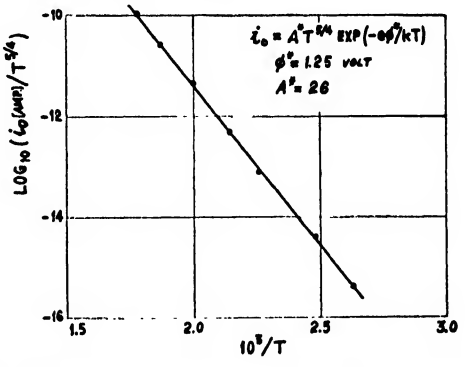


FIG. 7. Observed saturation current against temperature.

that predicted by Eq. (1) which is derived for an ideal uniform cathode. In the actual case, electrons emitted from the various parts of the cathode experience different retarding fields. The retarding potential which an electron experiences in going from the cathode to the anode is the difference between the height of the potential barrier at the anode surface and that near the patch from which the electron is emitted. The total retarding field current is the sum of the component currents from the various patches, each of which varies according to Eq. (1) when the proper value of retarding potential is used, and each saturates at a different anode potential. A plot of this total current against the applied retarding potential would show a slower approach to the saturation current than in the ideal case of a uniform cathode, and consequently an apparent shortage of slow electrons in the energy distribution of the emitted electrons. The effective work function of the non-uniform cathode can be defined as ϕ^* in the following equation:

$$J_0 = A^* T^{5/4} \exp(-e\phi^*/kT), \qquad (5)$$

where J_0 is the observed saturation current, and A^* is the effective A constant. Since in general the plot of $\log(J_0/T^{5/4})$ *vs.* $1/T$ for oxide cathodes can be represented by a straight line, this effective work function is equivalent to the work function for an ideal uniform

cathode which has the same thermionic emission as the actual cathode over the temperature range studied. The effective zero field V_0^* can be defined from the following equation:

$$V_0^* = \phi_{\text{anode}} - \phi^*. \tag{6}$$

It is determined from experiment by matching the experimental retarding field plot to the theoretical Schottky plot at both high retarding fields and at saturation. The zero field point on the theoretical curve marks the effective zero field, because when thus matched the theoretical curve represents the retarding field plot of an ideal cathode which has its work function equal to ϕ^*. The effective work function and effective zero field defined in Eqs. (5) and (6) are necessarily temperature dependent even if the work functions of the patches are independent of temperature, however the temperature dependences are only slight.

Experimental Result and Interpretation

The experimental study was made at temperatures between 442°K and 535°K. The maximum current density reached in these runs was below 0.1 μamp./cm^2. Experiment shows that at such low temperatures and low current densities no measurable change of the cathode state took place. The curves for increasing and for decreasing retarding potentials coincided very well. The changes in temperature during each run at one particular temperature were always within 0.2°C. Anode effect was found to be not present.

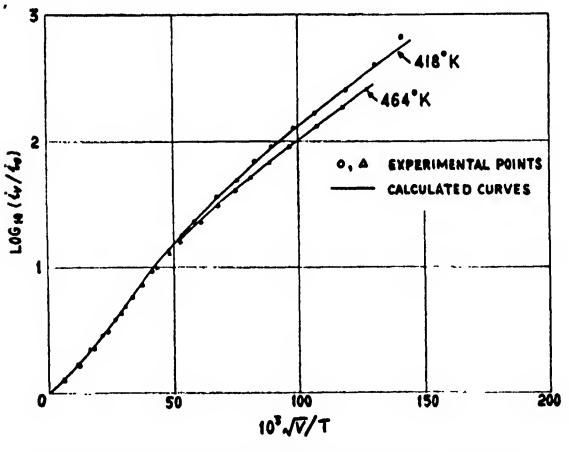


FIG. 8. Accelerating field plots for the same cathode as used for Figs. 4, 5, and 7.

The experimental curves were shown in Fig. 4. An enlarged drawing with all the curves matched together is shown in Fig. 5. It is seen that at retarding potentials above 0.2 volt, all the experimental curves coincide with the theoretical curve assuming no reflection. Therefore, at least for the high energy electrons, the energy distribution obeys Maxwell's law, and the electron temperature agrees with the cathode temperature. The departure of the experimental curves from the theoretical is seen to occur in a range of 0.3 volt. Even at the maximum current density of 0.1 μamp./cm, space charge distortion of the curves should not be detectable. The anode was found to be free from contaminations, so that the condition of uniform anode surface potential is approximately satisfied. Although the temperature of the cathode was not uniform along its whole length, it is believed that the temperature variation over the cathode under the collector was not more than 2 or 3°C, and the temperature variation over the entire cathode was within 30 or 40°C at the measuring temperatures we used. This temperature distribution along the cathode does not make the experimental curve deviate noticeably from the theoretical curve for a cathode with uniform temperature.

It remains now for us to examine the effects of first the non-uniformity of the cathode work function and second the reflection at the surfaces of the electrodes. The distribution of work function along the cathode being unknown, we use a crude model for discussion. We take our cathode as made up of patches that can be grouped into seven equal areas with work functions ranging in equal steps from ϕ to $\phi+0.15$ volt. As a first approximation, we neglect the distortion of the electron trajectories by the patch field, and sum up the component currents from the separate patches directly, using Eq. (1) with different retarding potentials for the different groups of patches. The calculated retarding field curve is shown in dot-dash line in Fig. 5 for the temperature of 500°K. Similarly calculated curves for temperatures from 450°K to 550°K show only slight deviation from each other. The experimental curves for the temperature range of 442°K to 535°K are also seen to be very close. The calculated curve fits the experimental curves quite well, and the use of a more elaborate model with a slightly larger range of work function variation would enable us to make a perfect match. It will be seen in the analysis of our accelerating field measurements that this value of work

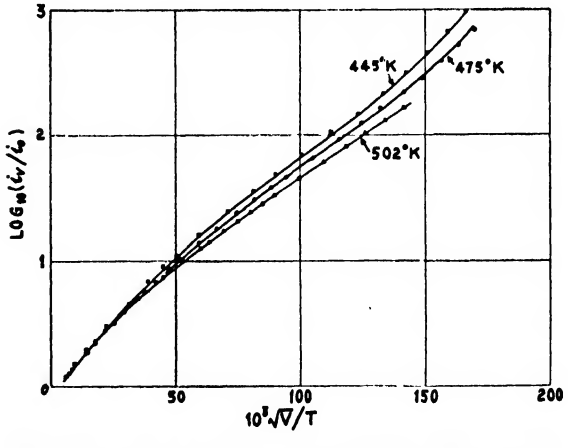


FIG. 9. Accelerating field plots for another oxide cathode; note similarity to plots in Fig. 8.

function variation agrees with the value we found there.¶

The Reflection Effect

As mentioned in the introduction, according to Nottingham's measurements on tungsten and thoriated tungsten filaments, a reflection coefficient as given by Eq. (2) with $k/\omega=4.5\times10^{-4}$ deg.$^{-1}$, should occur at both the cathode and the anode. Electrons that are reflected from the anode should go back to the cathode if the field is strictly radial. Since Nottingham used a filament wire of about 0.08 mm in diameter, a slight departure of the electrodes from being strictly coaxial would cause the electrons that are reflected from the anode to miss the cathode and go back to the anode again. Thus the anode current would not be affected appreciably by the reflection effect at the anode. However, with our tube, the cathode was 2 mm in diameter so that the electrons that are reflected from the anode had very good chance of landing on the cathode as they should in the ideal case. This reflection effect at the anode would reduce the anode current more at high retarding fields than at near zero field. Measurements yield information on the reflection effect only as the reflection near zero field differs relative to that at high retarding fields. Therefore the apparent reflection effect is greatly reduced when the reflection at the anode is included. The current expression for a uniform cathode when reflection effect is present is

$$J_v=(2e/h^3)(2\pi mkT)^{\frac{1}{2}}\exp(-e\phi/kT)\times\int_{-\infty}^{\infty}dp_y\int_{0;\,(2\text{ Mev}-(1-\theta^2)p_y^2)^{\frac{1}{2}}}^{\infty}\times(1-R)\exp\left(-\frac{p_x^2+p_y^2}{2mkT}\right)\frac{p_x}{m}dp_x, \quad (7)$$

where p_x and p_y are the normal and tangential components of the momentum of the electron at the cathode surface (the length of the cathode is along the z axis), and θ is as defined in Eq. (1). R is the reflection coefficient, and is given by Eq. (2) when reflection at the cathode alone is considered. When multiple reflections at both electrodes are considered, it is as follows:

$$1-R=\frac{(1-R_c)(1-R_a)}{1-R_cR_a}, \quad (8)$$

where $R_c=\exp(-p_x^2/2m\omega)$, $R_a=\exp(-P_x^2/2m\omega)$ and $P_x^2=p_x^2+p_y^2(1-\theta^2)-2$ Mev, P_x being the normal component of electron momentum at the anode surface. In Fig. 6 are shown the retarding field currents for different assumptions of reflection effects. It is seen there that while Curve B, where reflection effect at the cathode alone is considered, is markedly different from Curve A for zero reflection; Curve C, for which the multiple reflection effect is considered, has a form very nearly the same as Curve A. Therefore we see that the reflection effect, if it does exist, is not observable in our case. The possibility of the existence of an appreciable reflection effect at the electrode surfaces should not be excluded although we have found a satisfactory explanation of our retarding field measurements with the patch theory. A decisive check on this reflection effect would be to use a fine filament wire, in which case the reflection effect, if it does exist, would show up strongly.

Work Function and Zero Field

The thermionic emission against temperature curve is shown in Fig. 7. The currents used in the plot are the observed saturation currents so that the slope of the plot gives the effective work function. The points lie approximately on a straight line. The effective work function is found to be 1.25 volts and the A value around 26. The effective zero fields obtained from the matching of the experimental retarding field plots to the Schottky plot are shown below:

Temperature in °K	V_0^* in volts
535	3.17
501	3.19
468	3.21
442	3.22

This variation of the effective zero field is quite large and could be due to the changes in work functions of the patches and to the effect of the patch field distortion on the electron trajectories which we have neglected in our previous analysis. The anode work function obtained from the values of ϕ^* and V_0^* given above is around 4.4 volts. This is higher than the thermionic value of 4.07 volts obtained for tantalum by Dushman, Rowe, Ewald, and Kidner;[10] and the photoelectric value of 3.92 to 4.05 volts obtained by Cardwell.[11]

ACCELERATING FIELD STUDIES

Measurements

Emission currents were measured with accelerating potentials up to 7 kv and at temperatures below 500°K. The maximum electric field, calculated from the geometry of the electrodes, was 50,000 volt/cm at the cathode surface. The drop in potential across the coating was negligible according to Mahlman's measurements on the conductivity of oxide coating prepared in the same way as in this experiment. The emission

¶ It should be emphasized that what we mean here by the work function of the patches are the potential barriers, in the presence of the field due to the neighboring patches, which an electron climbs over in traveling along the normal to the patch area. In our later use of Becker's hill and valley checkerboard approximation for accelerating field studies, the work functions of the patches are as defined by Becker, i.e., they are the values when each patch is considered as an infinite plane with no neighboring patches.

[10] Dushman, Rowe, Ewald, and Kidner, Phys. Rev. **25**, 338 (1925).

[11] A. L. Hughes and L. A. Dubridge, *Photoelectric Phenomena* (McGraw-Hill Book Company, Inc., New York, 1932).

current was always below 1 μamp. and therefore there was no heating up of the cathode when high voltage was applied between the electrodes. The current was steady on the application of high accelerating potentials, and the current at low field did not change before and after a run was made. The results for the same cathode as we used in the retarding field studies are shown in Fig. 8. Measurements on two more cathodes show similar current voltage characteristics. Results on one of them are shown in Fig. 9. The curves are seen to decrease in slope as the voltage is increased and remain almost linear for a range of moderate voltages and increase rather fast in slope as 3 kv is exceeded. The minimum slopes near the linear part of the curves in Fig. 8 are about 2.6 times the theoretical Schottky line if one takes the field at the cathode surface to be that calculated according to the dimension of the electrodes. The slopes for the curves at the two temperatures are about the same. According to patch theory, we would expect a deviation of the experimental curve from the Schottky line at low voltages and that deviation should go to zero at high voltages. There are two difficulties if we were to rely on the patch theory alone to explain our results: first the rapidly rising parts of the curves at high voltages cannot be explained; and second, since the curves did not approach the theoretical Schottky line even at 5 kv, the average linear dimensions of the patches would have to be smaller than 10^{-6} cm. It is conceivable that there be patches of that size or smaller, but their contribution to the emission current would not be important owing to their small areas, unless their work functions are greatly reduced by a very strong accelerating field, a point which we will take up immediately.

The Effect of the Roughness of the Cathode Surface

Owing to the difficulties with patch theory alone, it is desirable to look into the other possible reasons for this deviation. It is easy to see that the roughness of the cathode surface has an important effect on the actual fields at the different parts of the cathode. With the oxide cathode built up of crystals of sizes of from one micron to a few tens of microns, piled up on one another, the field variation along the cathode would be very large. And it is likely to have a considerable fraction of the cathode surface with fields ten times or more higher than the average field. At low voltages, this effect will not be important. The lowering of the potential barriers at low accelerating potentials will be small even for the sharp points where the fields are many times higher than the average field. The change in the logarithm of the emission from a small area of the cathode is proportional to the lowering of the potential barrier there. When the latter is small, the change in the emission current itself is also nearly proportional to it, and therefore to the square root of the field. Thus the net field effective in increasing the emission for the entire cathode is close to the average field. (The effective field would be equal to the average field at low accelerating potentials if the lowering of the potential barrier were proportional to the field itself and not to its square root.) At high accelerating potentials, the lowerings of the potential barriers are large, and the change in the emission current at any spot is then no longer proportional to the lowering of the barrier. Emission currents from places where the fields are strong are far more important than those from places where the fields are relatively weak, and the effective field for the entire cathode is then very different from the average field, and is close to the field at the strong field places. It is thus seen that the ratio of the effective field to the average field increases from unity at low accelerating potentials to higher and higher values as the accelerating potential is increased. It is not surprising thus to find that the experimental curves did not approach the theoretical curve even at the highest average field of 50,000 volt/cm as we have used. The explanation of the entire curve is as follows: At low applied accelerating potentials the effective field is approximately the average field and the initial large slope of the curve and the subsequent decrease in slope is mainly due to the patch effect. At higher accelerating potentials, the patch effect becomes smaller but the effective field to average field ratio becomes greater so that the slope of the curve no longer decreases. For voltages above 1 kv, the patch effect becomes almost negligible and the curve has an almost constant slope over a considerable range of voltage. This slope is 2.6

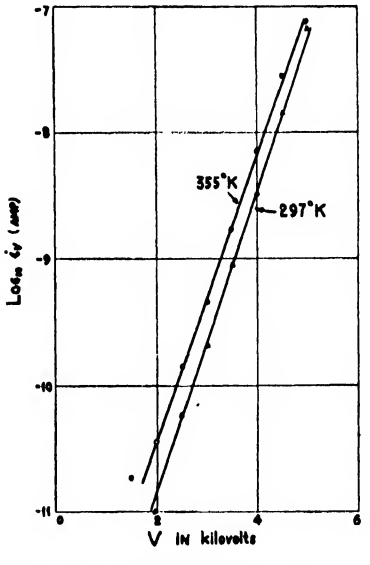


FIG. 10. Electron emission at near room temperatures under high accelerating voltages.

times higher than the slope calculated from the average field because the field effective in increasing the emission current is 2.6^2 times higher than the average field. For accelerating potentials above 5 kv, the slope of the curve increases rather fast because of the further in-

crease of effective field to average field ratio; and probably at some extremely sharp points, field emission was taking place. Based on this explanation, we can estimate the average linear dimension of the patches and the variation in work function of the patches using Becker's hill and valley checkerboard approximation.[12] Estimates have been made for two extreme cases: first, that the effective field to average field ratio is 2.6[2] throughout the whole range of accelerating potential up to 5 kv; and second, this ratio is unity for accelerating potentials below 1 kv and is 2.6[2] for potentials above that. The actual case should lie between these two extremes. The results of estimation are: the maximum work function variation (assuming sinusoidal variations from patch to patch) is 0.16 volt from average for both cases; and the average linear dimension of the patches are 2×10^{-5} cm and 2×10^{-4} cm for the first and the second cases respectively. The calculated curves for the first extreme case are shown in Fig. 8. This value of work function variation agrees with the value of 0.15 volt we have obtained in our analysis of the retarding field measurements. The linear dimensions of the patches of 2×10^{-5} to 2×10^{-4} cm are rather small compared to the sizes of the oxide crystals which run from 10^{-4} to 10^{-2} cm. It should be noted that in our analysis, perhaps it is around the smaller crystals that we have higher fields and therefore more important contributions to the observed current; and it is also likely that the patches correspond to crystal defects and not to the entire crystal faces.

Measurements at near room temperatures lend support to this view of the important effect of the roughness of the cathode surface. At high accelerating potentials, fluorescent spots appeared on the glass wall of the experimental tube indicating that the emission came from local spots of the cathode where the fields are very high. Curves in Fig. 10 show that the current is temperature sensitive suggesting that at least part of it is thermionic emission from local sharp points where the potential barriers are lowered to perhaps two- or three-tenths of an electron volt by the extremely high fields there.

It is suggested that further accelerating field measurements be done with single crystals or with cataphoretically deposited cathodes whose surfaces are far more smooth than the ordinary cathodes with sprayed-on coatings.

The author wishes to express his gratitude to Professor W. B. Nottingham for the suggestion of this work, his constant interest in it and his many helpful suggestions.

[12] I. A. Becker, Rev. Mod. Phys. 7, 95 (1935).

Augmented Flames in Half-Open Tubes[1,2]

MARJORIE W. EVANS, MILTON D. SCHEER, LOUIS J. SCHOEN, AND EMMY L. MILLER
Research Division, New York University, University Heights, New York City
(Received July 18, 1949)

High velocity flames in half-open tubes produced by placing a grid augmenter in the path of an advancing flame are described. High speed schlieren motion pictures and instantaneous pressure measurements are used to study flame structure and accompanying flow in the tube. It is found that the grid augmenter causes a mixing of burning and unburned gases, resulting in an increase in reaction surface. As a result the total amount of reaction per unit time per unit volume is increased, thus leading to the observed high velocities.

INTRODUCTION

THE velocity of propagation of flames burning in gas is determined by the normal combustion velocity (which is dependent upon thermal conductivity, diffusion, and the detailed reaction kinetics), and by the flow processes in the system. A hydrocarbon-air mixture ignited by a spark at the closed end of a half-open tube (i.e., a tube closed at one end and open at the other), generally has a maximum flame front velocity of about 100 feet per second.[2] Certain flames such as those in pulse jet engines have been observed to move with a considerably greater velocity. Flames of high velocity, in some cases approaching 1000 feet per second, have been produced in this laboratory by placing in the path of the advancing flame a grid augmenter consisting of a sheet of metal containing holes with trailing edges. Such a sharp increase in velocity after the flame front has passed through a twenty-hole grid may be seen in Fig. 1, which shows the flame velocity recorded in an experiment in which the grid was placed two feet from the closed end.

This paper reports investigations into the nature of the high velocity or augmented flame produced by the grid augmenter, utilizing schlieren techniques in conjunction with a high speed motion picture camera, and pressure measurements.

[1] This research was conducted under the auspices of Project Squid, sponsored by the ONR under Contract N6ori-11, Task Order II.

[2] This material is a further development of work first reported at the Third Symposium on Combustion, Flame, and Explosion Phenomena, University of Wisconsin, Madison, Wisconsin, September, 1948, in a paper entitled, "A study of high velocity flames developed by grids in tubes," Evans, Scheer, Schoen, and Miller.

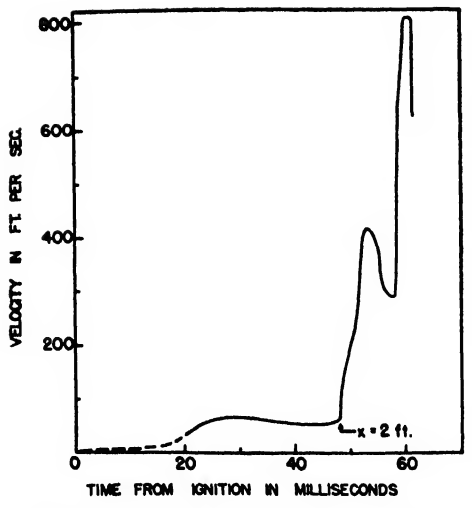


FIG. 1. Velocity-time curve for a flame front in a tube in which a twenty-hole grid augmenter was placed two feet from the closed end.

SCHLIEREN AND PRESSURE MEASUREMENTS

The flame tube was six feet in length with an inner square cross section approximately 3.5 inches on a side. The grid augmenter in these experiments had three rectangular nozzles 0.3 square inch in area with trailing edges 0.3 inch long. The grid was placed 3.1 inches from the spark igniter, which was supplied by a high frequency induction coil. The tube was filled with a 6.6 percent by volume propane-air mixture.

The optical arrangement consisted of a double-mirror knife-edge schlieren system with parabolic mirrors sixteen inches in diameter, having sixty-inch focal lengths. The parallel beam of light between the two mirrors passed through the plate glass side-walls of the flame tube. The light source was a Western Union 100-watt zirconium arc lamp, and a 16 mm Fastax camera, with a speed booster, was used at a speed of 4000–9000 frames per second to record the gas motions in the tube. In order to follow the motion of a particular layer of gas during the flow which accompanies combustion, observations were made of the positions of local thermal gradients produced by stretching across the tube a nichrome wire heated electrically to a temperature well below the ignition point.

Pressure measurements were made with the frequency modulated condenser type pressure gauge developed in this laboratory by J. H. Hett and R. W. King, Jr.[3] The time of ignition, t_{ig}, was correlated with the occurrence of the pressure pulse and with events as recorded in the schlieren motion pictures by means of the circuit shown in Fig. 2. This circuit utilized the light emitted by the igniting spark, as picked up by a photo-cell, to trigger indicating lights which registered on the motion picture film and on the film recording the pressure trace.

Figure 3 is a reproduction of nine selected frames from a schlieren motion picture of the flame as it passes through the grid augmenter. Apparently, in this case only two holes were effective. The flame before reaching the grid shows several large cells and has a relatively smooth surface. These cells strongly resemble those recently observed by Markstein in flames held stationary in tubes by adjustment of the flow velocity.[4] The augmented flame, on the other hand, shows a much more complex structure with fine-grained eddy motion. As it progresses down the tube it assumes a more homogeneous structure. The faint patterns showing gas flow through the grid nozzles in pictures *B*, *C*, and *D*, of Fig. 3 are caused by gas which has been heated by the flame.

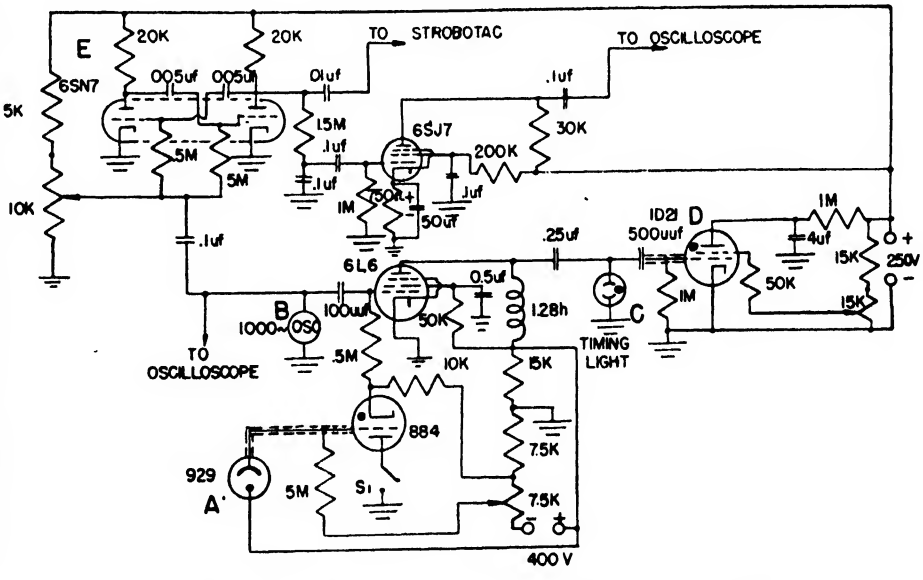


FIG. 2. A circuit for the simultaneous indication of zero time on the schlieren motion picture and pressure recording film. The photocell *A*, picking up light from the igniting spark, triggers a thyratron tube which allows the oscillator *B* to energize the timing light *C* in the Fastax camera, thereby providing a zero time mark t_{ig}, as well as timing marks at 1000 cycles per second for the schlieren motion pictures. *D* is a gas tube which flashes when the photo-cell begins to conduct, thus providing a zero time mark on the pressure recording film. *E* is a frequency divider which divides the 1000-cycle output by 4, thereby furnishing 250 cycles per second to a strobotac which provides timing marks for the pressure recording film. The oscilloscopes were used for the calibration of the oscillator and the frequency divider in conjunction with a 1000 cycle tuning fork oscillator as a standard.

[3] J. H. Hett and R. W. King, Jr., "A frequency modulation pressure recording system," Rev. Sci. Inst. (to be published).
[4] G. Markstein, J. Chem. Phys. 17, 428 (1949).

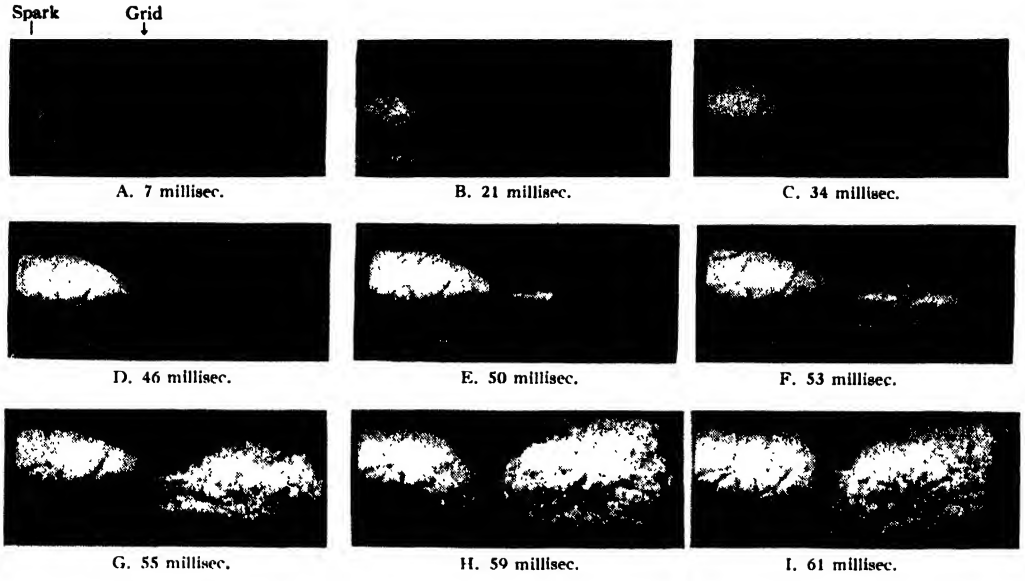


FIG. 3. Schlieren motion pictures taken at 4400 frames per second of flame moving through a three-hole grid augmenter.

Figure 4 shows schlieren pictures of the progress down the tube of a layer of gas in front of the flame. In these early pictures is seen a spark indicator, in series with the igniting spark, placed in the field of view to indicate the time of ignition. This method was later replaced with the timer circuit mentioned above. The position of the heated wire used to produce the thermal gradient is shown in *A*. In subsequent pictures the gradient is observed moving toward the open end. The flame comes into the field of view in *C*, and overtakes the gradient in *E*. The data plotted in Fig. 5 were obtained from measurements made from pictures such as these.

Figure 5 is a plot of position *versus* time for layers of gas as the augmented flame expands and moves down the tube. The solid line represents the motion of the flame front, the dashed lines that of layers of gas which were labeled with thermal gradients. The plot shows that all gas between the grid and a point ten inches from the grid at zero time is enclosed by and becomes part of the augmented flame. This volume is termed the augmentation region. The gas which is initially beyond this ten-inch region is pushed ahead of the flame during the expansion which occurs as a result of the combustion reaction, and is not burned.

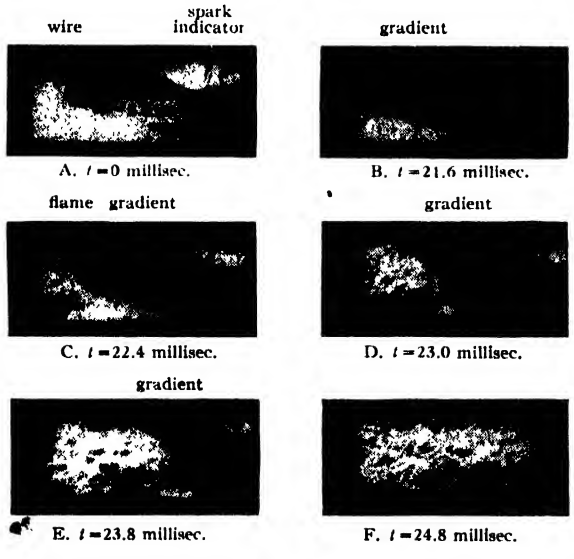


FIG. 4. Schlieren motion pictures taken at 4300 frames per second showing a thermal gradient moving down the tube ahead of an augmented flame.

While the position-time curves for unaugmented flame fronts vary radically from experiment to experiment, the characteristics of the augmented flames are quite reproducible.[2] Under identical experimental conditions the position-time points for several flame fronts fall on the same curve. The lateral growth of the flame in the augmented region is likewise reproducible. A plot of the growth of the augmented flame in the y direction (perpendicular to the axis of the tube) at a point 2.5 inches from the grid augmenter is shown in Fig. 6. It is at this point that the developing augmented flame first reaches the sides of the tube.

Pressure measurements shown in Fig. 7 were made with the gauge placed at points *B*, *C*, and *D*, in the tube as indicated in the figure. The time t_p in Fig. 6, when the flame reaches the sides coincides with the beginning of the first pressure pulse at position *B*. This suggests that the origin of the pressure pulse is in the region near *B*. Pressure measurements in a tube without grid showed only a small rise over atmospheric pressure, the peak pressures being approximately a factor of ten smaller than those shown in Fig. 7. When the tube with grid was filled with propane-air mixture, simultaneous pressure measurements at *B* and in the initiating chamber (i.e., between spark and grid) gave two essentially

identical curves, the method of timing not being accurate enough to indicate the difference in the time of occurrence of the pulse at the two points. When only the initiating chamber was filled with gas, the reaction produced only a small pressure rise. This confirms the conclusion that the pressure pulse is generated near position B, in the augmentation region.

Observation of the schlieren motion pictures shows that at time t_p when the flame reaches the sides of the tube the region 2 to 2.5 inches downstream from the grid augmenter is the center of an expansion of the burning gases. The observation together with the facts cited above indicates that the augmented flame is produced in the augmentation region by a mixing process between t_g (when the flame emerges from the grid augmenter) and t_p. After time t_p the dominant phenomenon is rapid burning and expansion of burning gases with the consequent rapid motion of the flame toward the open end of the tube. It has been previously noted[2] that streak photographs of the augmented flame show that the flame apparently first appeared several inches below the grid, and proceeded to burn simultaneously toward the open end and back toward the grid. This can now be explained, in view of the above observations, by assuming that burning vigorous enough to produce sufficient illumination to record a trace does not occur until time t_p, with the vigorous burning beginning at the center of expansion, 2 to 2.5 inches downstream from the grid.

DISCUSSION

It is clear that the high flame front velocity of the augmented flame is to be attributed to a rapid expansion during the burning of the gas in the augmentation region. In the normal unaugmented flame in the same tube somewhat less gas is burned, but over a longer period of time. This greater rate of burning of the augmented flame is due to an increased burning area or reaction surface. When the burning or burned gas from the initiating chamber flows through the nozzles into the stationary gas a mixing process occurs so that burning surfaces are interspersed with unburned gas. As a result there is more burning area per unit volume than in the normal flame, which presents a relatively smooth surface with a few large cells.

The details of the mixing are not clear. The two-dimensional problem considering the spread of a mixing region in the case of the turbulent mixing of a stream of fluid (of uniform velocity at the mouth of the nozzle) with a fluid at rest has been analyzed by others.[5] The

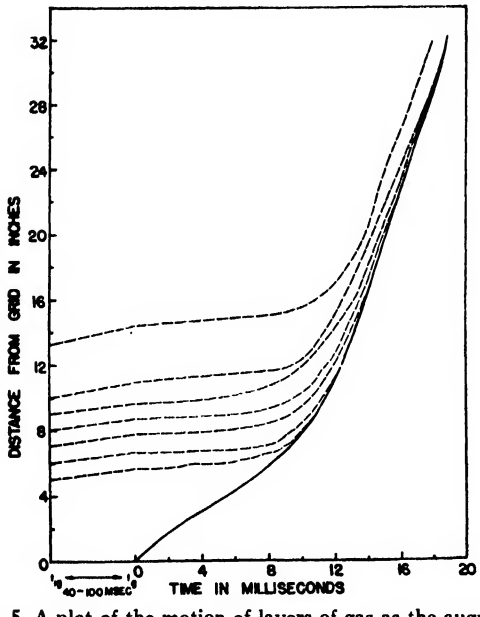


FIG. 5. A plot of the motion of layers of gas as the augmented flame grows and moves down the tube. The solid line represents the flame front, the dashed lines layers of unburned gas ahead of the flame. t_{ig} is the time of ignition, t_g the time of appearance of the flame at the grid. Measurements were made from motion pictures taken at 5000–7500 frames per second.

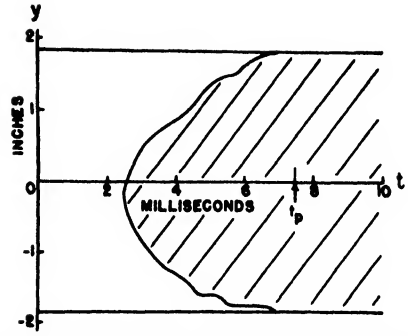


FIG. 6. A plot of the rate of growth, in the y direction, of the mixing region, at a point 2.5 inches below the grid. $y=0$ is the axial line of the tube. Time is measured from the appearance of the flame at the grid. t_p is the time of the beginning of the first pressure pulse shown in B of Fig. 7.

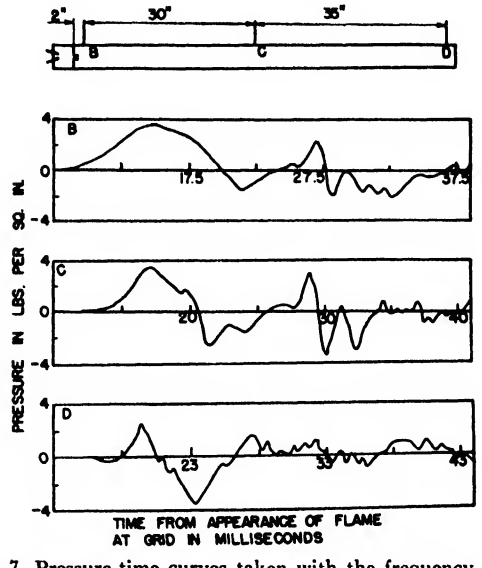


FIG. 7. Pressure-time curves taken with the frequency modulated condenser type pressure gauge at the points B, C, and D, in the augmented flame tube. The flame tube, with spark and grid augmenter, is shown diagrammatically.

[5] *Modern Developments in Fluid Dynamics*, edited by S. Goldstein (Oxford University Press, New York, 1943), Vol. II, p. 597.

TABLE I. Comparison of average pressure and specific impulse of a small pulse jet engine using gasoline as fuel, and the augmented flame tube using a propane-air mixture (6.6 percent by volume) as fuel.

	Small pulse jet	Augmented flame tube
Average pressure (lb./sq.in.)	3.3	1.1
Specific impulse (lb. thrust/lb.fuel/sec.)	760	410

present situation is more complicated in at least four respects, *viz.*, 1) the flow field is not two-dimensional, 2) the velocity of the discharging jet at the mouth of the nozzle is not constant, 3) three or more nozzles are in juxtaposition, and 4) a combustion reaction at an elevated temperature is occurring in the jet and at the boundaries. The qualitative observation may be made from Fig. 3 that the core of fluid in which laminar flow occurs disappears within two or three nozzle diameters downstream, after which the mixing region occupies the whole jet. One can speculate that regions of vorticity and turbulence spread from each jet and ultimately interfere with one another creating additional unstable swirling type flow of an even more complex nature. In this process large vortices may perhaps break up into smaller ones during the lateral growth of the mixing region.

It is of interest to inquire as to whether the augmented burning described here resembles the burning observed in jet engines, particularly with respect to the rate of burning. This is conveniently done by comparing the average pressure in a typical small-scale pulse jet, which uses gasoline as fuel, with the average pressure in the flame tube. In order to carry out this comparison it is assumed that the flame tube could be fired at a frequency which would correspond to a period comprised of the time for the pressure to rise from atmospheric to a peak, back through a minimum, and then to atmospheric again, and that such repetitive firing would not materially alter the pressure record, shown in Fig. 7. Such an assumption is suggested in part by the fact that a similar sequence of pressure changes occurs in a combustion chamber of a pulse jet during a cycle of operation. At any rate, following this hypothesis, if the curve B of Fig. 7 is integrated from 7.5 to 23.9 milliseconds and the result divided by the elapsed time the average pressure in the flame tube is obtained, i.e.,

$$P_{Av}=\left(\int_{7.5}^{23.9} P(t)dt\right)\Big/ 16.4.$$

The specific impulse is calculated on the assumption that the augmentation region is filled with propane-air mixture at 16.4 millisecond intervals. Inasmuch as the heats released per unit mass for gasoline and propane differ by only 5 percent and the calculations are only approximate, no correction was made for the difference. Table I lists the results in terms of average pressure and specific impulse, and the quantities are found to be of the same order of magnitude for the small pulse jet and the augmented flame tube.

CONCLUSION

Despite an incomplete understanding of the mixing process in the flame tube, it is possible to draw some conclusions with regard to the effect of swirling or eddy motion on combustion processes. Such motion is an important factor in determining the rate at which a given fuel-air mixture can be burned, its function being, at least under the conditions of these experiments, to increase the total reaction surface, thus increasing the amount of mixture reacting per unit time per unit volume.

In situations where sufficient fuel and air are available an increase in the intensity of the swirling and mixing may make possible a greater rate of energy release in the same reaction chamber volume. Conversely, under these conditions for a given desired rate of energy release a smaller volume of combustion chamber may be used. These possibilities could conceivably be of interest in the design and operation of industrial furnaces and internal combustion engines as well as jet engines.

The authors take pleasure in acknowledging their indebtedness to the late Professor J. K. L. MacDonald and to Dr. J. H. Hett, Professor G. E. Hudson, and Messrs. R. W. King, Jr., and Simeon Braunstein for help received during these studies.

Static Magnetic Storage and Delay Line*,**

AN WANG AND WAY DONG WOO
Computation Laboratory, Harvard University, Cambridge, Massachusetts
(Received July 27, 1949)

Magnetic cores with a rectangular hysteresis loop are used in a storage system which requires no mechanical motion and is permanent. The binary digit "1" is stored as a positive residual flux, and the binary digit "0" as a residual flux in the opposite direction. When a negative probing field is applied to the core, a large voltage is induced in another winding if the digit stored has been a "1," and very small voltage if it has been a "0." The induced voltage in the former case is large enough to magnetize another core of identical construction. Binary digits can thus be transferred from one core into another. Many cores are arranged in tandem to form an information delay line. Binary digits can be advanced along the line step by step. The present upper limit of the speed of propagation is about 35,000 digits per second, and there is no lower limit.

I. INTRODUCTION

ONE of the most important research problems in the field of large-scale digital calculating machines is that of information storage. Electronic trigger pair tubes, cathode-ray tubes, acoustic delay lines, rotating drums, or tapes are used in present machines. Electronic trigger pair storage is, of course, the most versatile, but it is extremely expensive and its reliability poor. Cathode-ray tube and acoustic delay line storage systems need constant regeneration to preserve the information stored. All three systems lose the information if there is interruption of power. Magnetic drum or tape storage preserves information indefinitely, but it involves a mechanical system, and the operating speed is limited by mechanical considerations. Both in acoustic delay lines and rotating magnetic drums, the operation is necessarily synchronous. The speed depends upon the physical nature of the system and cannot be changed at will during operation.

This paper describes a storage system which operates like electronic trigger pairs but does not use vacuum tubes other than a relatively small number for control purposes.

A binary digit which takes the value of "0" or "1" is stored in a ferromagnetic core in the form of positive or negative residual magnetism, depending on the direction of the last magnetizing force. As shown in Fig. 1, a binary digit is represented by the point "0" or "1" on the hysteresis curve of the core. To store a binary digit, only the application of a magnetizing pulse of proper polarity is necessary.

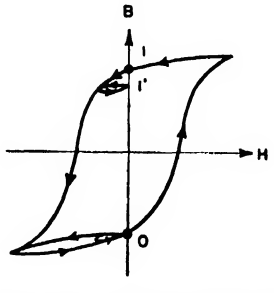


FIG. 1. Hysteresis loop.

The reading out of the stored information presents a more difficult problem. For a small negative pulse of magnetizing force H, applied at point "0," there results less change in the flux density B than for a similar negative pulse applied at point "1." Thus the relative amplitude of the induced voltage indicates what information has been stored in the material. However, this is true only for the first-read-out pulse, since after "0" information has been read out, the material returns to the point "0," whereas after "1" information has been read out, the material goes to a new point "1'" instead of returning to "1." Further negative pulses of H produce very nearly equal changes of B for starting positions "0" and "1'." Thus, whereas small read-out pulses do not actually destroy the positive or negative residual magnetism representing the stored information, they make reading out by like pulses impossible after the first.

When a large negative read-out pulse of H is impressed, the information is destroyed. The state of the material returns to position "0" in the B–H curve, no matter what information was previously stored. However, the secondary induced voltage is much larger when B changes from "1" to "0" than when B changes from "0" to "0." The ratio of these voltages was found to be as high as 30:1 for suitable magnetic materials. The large induced voltage can be made to drive another core from state "0" to "1." Thus one can construct (1) a storage unit in which digits can be stored and read-out continuously, and (2) an information delay line in which a series of binary digits can be propagated at

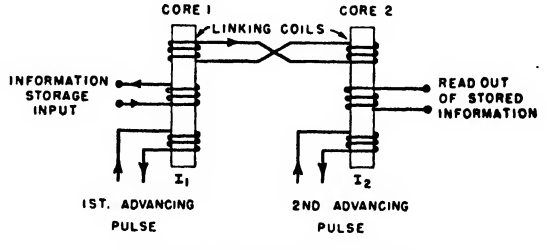


FIG. 2. Two-core storage unit.

* Work done under Contract W19-122-ac-24 between Harvard University and the United States Air Force.
** Presented at the Cambridge Meeting of the American Physical Society (June, 1949).

any desirable speed up to about thirty thousand information pulses per second.

II. A SINGLE-INFORMATION STORAGE UNIT

Since the large secondary induced voltage is sufficient to drive another core of similar size to its saturation, a means is available to transfer the information.[1] Such a storage unit is shown in Fig. 2. Core 1 and Core 2 are wound of strips of material with a nearly rectangular hysteresis loop to insure that the states of residual magnetism, 1 and 0, are well defined and stable. Assume

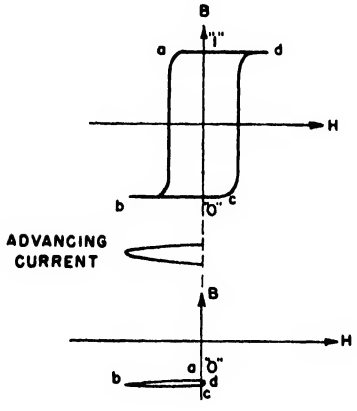


FIG. 3. Flux changing path.

that the 1 state of residual magnetism corresponds to flux in the cores in the upward direction, and the 0 state to flux in the downward direction. Then, whenever a 1 is stored in Core 1 through the information input coil, the first current pulse I_1 will change its state from 1 to 0. This generates a large induced voltage in the linking coil and drives Core 2 to state "1." After that the second current pulse I_2 shifts this state "1" back to Core 1 again. Thus the state "1" is transferred between Core 1 and Core 2 when the current pulses I_1 and I_2 are applied alternately. However, if both cores are originally in the "0" state, the I_1 pulse cannot change much of the flux of Core 1. The linking current is very small, and Core 2 remains in the "0" state. The alternate application of I_1 and I_2, which will be called advancing pulses from now on, therefore will not produce any change of the states of the cores. This constitutes a single-information storage, in which a binary digit can be stored indefinitely and read out at any given time by the application of a pair of advancing pulses.

Physically this is possible due to the existence of two limit cycles of the hysteresis loop when the two alternate advancing pulses are applied. As shown in Fig. 3, the upper loop represents the cyclic change of magnetization of cores when digit "1" is stored. The lower small loop represents the cyclic change of magnetization of cores when digit "0" is stored. In either case an advancing pulse changes the magnetization of one core along the curve *a* to *b* to *c* while that of the other core, through the linking coils, changes from *c* to *d* to *a*. Depending on the initial state of magnetization of the first core, the hysteresis loop follows either the upper or the lower curve.

III. PROPAGATION OF INFORMATION ALONG A DELAY LINE

The requirement for a magnetic delay line is that it should be able to store a series of information pulses and deliver them back at a later time. A number of magnetic cores are needed. These should be so constructed as to work in the following manner: When the first information pulse has been stored in the first magnetic core, an advancing pulse is applied to this core to transfer the information to the next core. This is necessary in order to make the first core ready to receive and store the second-information pulse. Next, these two signals are advanced further to make room for the storage of the third. This process of advancing and storing continues to the end of the line, where the information signal is sent out. If, after the line is loaded with information, no further advancing pulses are applied, the information will stay in the cores indefinitely. When advancing pulses are once again applied, the stored information is discharged. If the output of the delay line is connected to its input, a closed ring results. Information can be kept circulating around the ring continuously by the advancing pulses, and can be reproduced at any time. The same is true for magnetic recordings on a rotating storage drum, except that in the latter case the mechanical rotation of the drum is used to carry the information along, whereas in the present scheme advancing pulses are used to propagate the information.

Since all the magnetic cores in such a delay line are connected in tandem through the linking coils, two possible effects must be prevented before the desired information transfer can be made. First, an advancing pulse should advance a digit only in the forward direction; any backward flow of information is to be avoided. Second, an advancing pulse should advance a digit to the next core only; any effect of the digit on cores beyond this would interfere with previous digits. Of course it is imperative for successful operation that the signal should not fade along the line.

IV. A DELAY LINE OF THREE CORES PER BINARY DIGIT

A circuit has been developed which is based on the fundamental requirements of a delay line, discussed in Section III. The connections are shown in Fig. 4. Three cores are necessary for the storage of each binary digit. Three such cores together will therefore be referred to as one storage unit. Three successive pulses are required to advance the information from one storage unit to the next. These pulses, respectively, advance the signal from the first core to the second core,

[1] Chapter IV, Progress Report No. 2, November 10, 1948; the Computation Laboratory, Harvard University, under Contract W19-122-ac-24 with the U. S. Air Force.

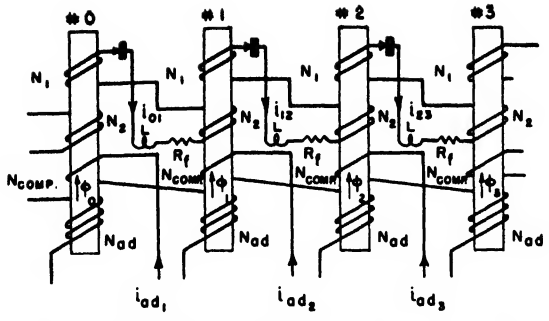


FIG. 4. Delay line structure of three cores per digit.

from the second core to the third core, and then from the third core to the first core of the next storage unit.

The operation can briefly be summarized as follows: All advancing pulses are connected so as to drive all the cores to negative saturation. When there is no information in the line, all the cores have negative residual magnetism and are thus in the "0" state. When information is stored, every third core may be positively or negatively magnetized and thus be in either state "1" or "0." Consider the case of a "0" being stored in Core 1. All advancing pulses are so connected so as to drive the cores to negative saturation. Since the residual magnetism on Core 1 is already negative and the core has essentially a square hysteresis loop, only a very small change of flux results when Advancing Pulse #1 is applied. There is no effect on other cores. The next core remains negatively magnetized. One can consider that the digit "0" information has been transferred to Core 2. On the other hand, if a "1" is stored in Core 1, it is positivcly magnetized. The application of i_{ad1} drives it to negative saturation. This large change of flux in Core 1 induces a large electromotive force in coils N_1 and N_2 of that core, causing i_{01} and i_{12} to flow. The current i_{12} drives Core 2 from negative magnetization to positive saturation.

The current i_{12} satisfies the following equation

$$N_1(d\phi_1/dt)+N_2(d\phi_2/dt)=[R_f+L(d/dt)]i_{12}, \quad (1)$$

where R_f represents the total resistance of the link circuit in the forward direction of i_{12} and L the leakage inductance of the linking coil. Integrating Eq. (1),

$$N_1\Delta\phi_1+N_2\Delta\phi_2=R_f\int i_{12}dt+L\Delta i_{12}.$$

Note that i_{12} is always a positive quantity due to the presence of the rectifier. Δi_{12} vanishes as the fluxes of Cores 1 and 2 reach a constant value. We get

$$\Delta\phi_2=-(N_1/N_2)\Delta\phi_1+R_f\int i_{12}dt. \quad (2)$$

The last term in Eq. 2 is always a positive quantity. If Cores 1 and 2 have the same dimensions, the flux change of ϕ_2 is not complete unless N_1 is greater than N_2. The optimum ratio of N_1 to N_2 is found to be close to 2:1.[2] The resultant change of flux of Core 2 has no effect on Core 3 because the linking current i_{23} is blocked by the rectifier in the circuit. This satisfies the requirement that information may not advance to cores beyond the next one. The current i_{01} is not desirable because it has the tendency to transfer information backward from Core 1 to Core 0. This current must be balanced by the use of a compensating winding, so that the net magnetizing force is always negative in Core 0 to maintain its negative saturation when information is being transferred from Core 1 to Core 2. Since this process of transfer affects only the three cores 0, 1, and 2, another digit in Core 4 can be transferred at the same time without interference in any other cores except 3, 4, and 5.

A static magnetic storage system is then obtained in which three cores are needed to store each binary digit. Information can be stored in the cores indefinitely and can be stepped along the line in the forward direction by the application of advancing pulses. Three advancing pulses are necessary to advance the information by a single storage unit.

Figure 5 shows a breadboard connection of such a delay line for the storage of four binary digits. The information can be advanced through each storage

FIG. 5. Four binary digits delay line—three cores per digit.

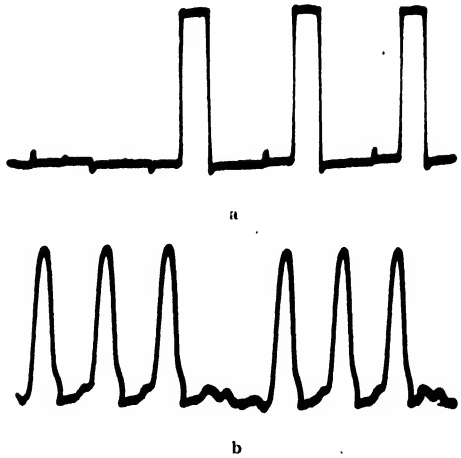


FIG. 6. Flux *vs.* time curve—three cores per digit. a. Information rate=3 kc. Information 0111. b. Information rate=30 kc. Information 1110.

[2] Chapter V, Progress Report No. 3, February 10, 1949; the Computation Laboratory, Harvard University, under Contract W19-122-ac-24 with the U. S. Air Force.

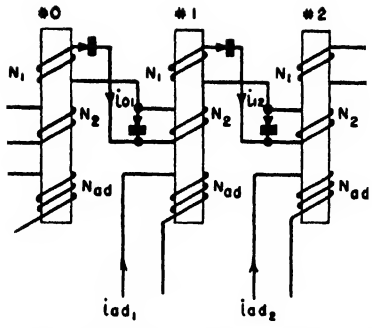


FIG. 7. Delay line structure of two cores per digit.

unit of three cores as fast as thirty thousand digits per second. Oscillographs showing flux variation of a core rate of 3000 and 30,000 digits per second are given in Fig. 6.

Small selenium rectifiers are used in the linking circuit. The rectifiers can be replaced by saturable reactors which offer a high reactance for current in one direction and a low reactance for current in the other direction. But since a reactor is not a highly dissipative element, its use prevents the linking current from dying down fast. This greatly reduces the maximum possible rate of advancing.

An analytical investigation of the sotrage unit in which rectifiers are used in the link circuits is given in reference 2.

V. A DELAY LINE OF TWO CORES PER BINARY DIGIT

As shown in the last section (Fig. 4), the compensating winding is necessary to counteract the effect of the current i_{01} flowing in the coil N_1 of Core 0. If this current can be eliminated, clearly no compensating winding is necessary. It can be done by the introduction of another rectifier,[3] as shown in Fig. 7. In the drawing the leakage inductance and circuit resistance have been omitted in the linking circuit for simplicity. The shunting of a rectifier across the coils N_2 is in such a direction that the current i_{12} is unaffected and produces the useful effect of transferring information from Core 1 to Core 2. But the current i_{01} is greatly reduced by the rectifier which short-circuits the electromotive force generated across N_2 of Core 1. The information then does not flow backward. Since only two cores are now involved in transferring the information, the next two cores can be used for another digit. It is then possible to use only two cores per binary digit.

Figure 8 shows an assembly of a delay line of this design to hold ten digits. Only two advancing pulses are needed to advance the information along the line. A few patterns of the flux variation in cores of this type of delay line are shown in Fig. 9, the repetition rate being 2.5 kc and 25 kc.

The only disadvantage of this connection is that the current flowing in coil N_2 of Core 1 is larger than before. This necessitates a more powerful advancing pulse to change the flux of Core 1.

Another possible arrangement is to apply a series voltage in the link 0–1 when i_{adv} is applied to Core 1 (Fig. 10). This series voltage is opposite to, and slightly greater than, the voltage generated in coil N_2 of Core 1. The transfer of information from Core 1 to Core 2 is made still easier. This scheme has also proved to be successful.

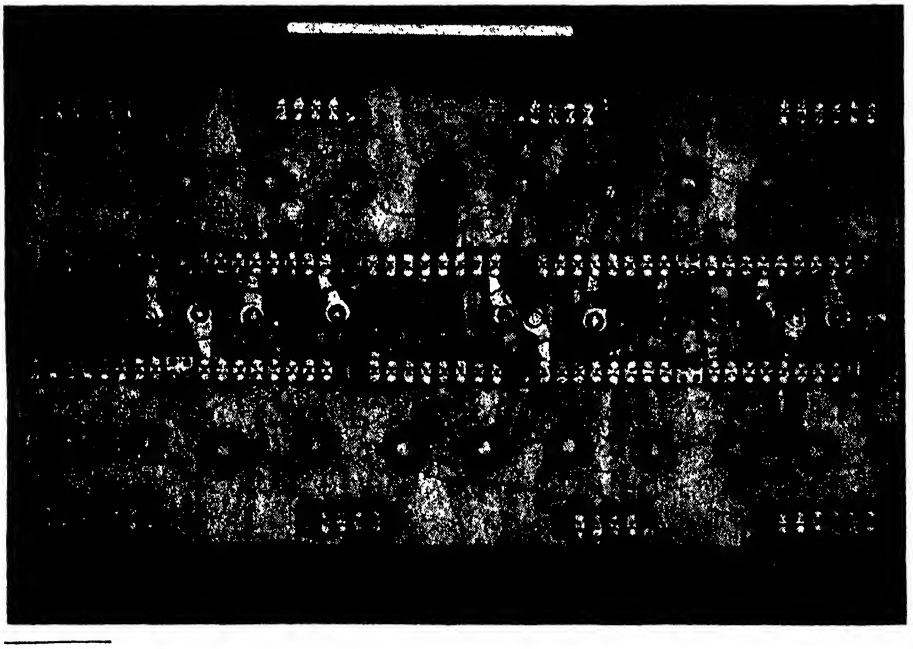


FIG. 8. Ten binary digits delay line—two cores per digit.

[3] Chapter V, Progress Report No. 4, May 10, 1949; the Computation Laboratory, Harvard University, under Contract W19-122-ac-24 with the U. S. Air Force.

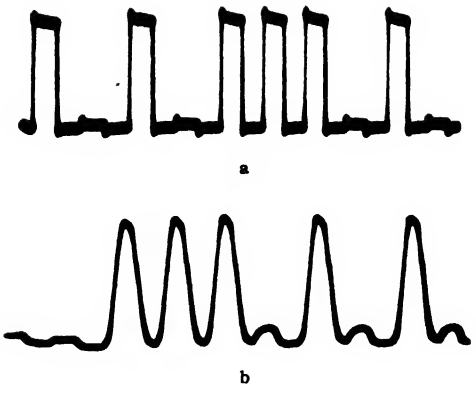


FIG. 9. Flux *vs.* time curve—two cores per digit. a. Information rate = 2.5 kc. Information 1010111010. b. Information rate = 25 kc. Information 0011101010.

VI. PROPERTIES OF MAGNETIC CORE MATERIAL

As shown in Fig. 11, the ideal hysteresis characteristics of the magnetic storage and delay line are a perfectly rectangular hysteresis loop and a very low coercive force.

In actual practice, unfortunately, such characteristics cannot be realized. Figure 12 shows a typical hysteresis curve for one of the cores manufactured by the Allegheny-Ludlum Steel Corporation of wound strips of Deltamax. The hysteresis loop is close to the ideal; nevertheless, the saturated region is not absolutely of zero slope, and the sharply rising portion is not quite vertical. In addition, the corners are somewhat rounded. But delay lines using such cores have been constructed and operate reliably. The criteria of how good the hysteresis loop is must have been discussed in reference 2.

As the frequency of operation becomes higher, the hysteresis and eddy current losses become appreciable. The heat generated is considerable at the highest frequency of operation. Moreover, the eddy current acts like a short-circuited secondary around the core. The presence of the eddy current requires a larger advancing current to drive the core from one polarity to the other.

At an operating frequency of 50 kc, every core alternates its flux between two saturation values in 10 microseconds. Assume that the flux changes linearly. This is a fairly accurate assumption, as seen from an oscillograph of flux variation. The cores used are made of a strip of one-mil thickness, $\frac{1}{8}$ inch wide, wound into a single-turn core of $\frac{1}{2}$-inch diameter. The hysteresis loss is estimated to be of the order of 7 milliwatt/core, the eddy current loss 25 milliwatt/core, the electrical resistivity of the core being taken as 50 microhm centimeters. For such a core the equivalent demagnetizing ampereturns can be calculated to be about 1.2. It is also equivalent to an increase of coercive force of 0.37 oersted. Apart from the upper frequency limits imposed by these losses, the fundamental operation of the storage and delay line is independent of the frequency.

VII. APPLICATIONS

It may be helpful to point out the unique features of this form of storage. No power is needed to maintain the storage of information. No mechanical movement is needed in recording and pick-up. The speed of recording, transfer, and read out has no lower limit and can be varied to an upper limit which is sufficiently high to be very useful. At the present stage, this maximum rate is of the order of 50 kc. These features represent improvements over the characteristics of the storage devices mentioned in the introduction.

Possible applications include the following:

1. *Information storage.* Individual binary digits can be stored in pairs of cores, as discussed in Section II. If a series of digits is to be stored, the information can be sent into a delay line. After the whole series of digits is in the line, stopping the flow of advancing pulses suffices to store the information indefinitely in the line. It can be read out at any time at the end of the line by applying the advancing pulses again. Development work is being done at the Computation Laboratory of Harvard University to construct a line holding 40 binary digits in a plug-in component container unit which will be no more than two inches in diameter and six inches in length.

2. *Input and output link for high speed computers.* This form of delay line provides an excellent medium for transfer of information between systems of different digital rate. Data and commands can be put into the line manually and then read out to the machine at its operating speed. Again, results from the machine can be fed at high speed into the delay line, from which later a set of slow advancing pulses can step the results out at a speed suitable, say, to operate a typewriter directly.

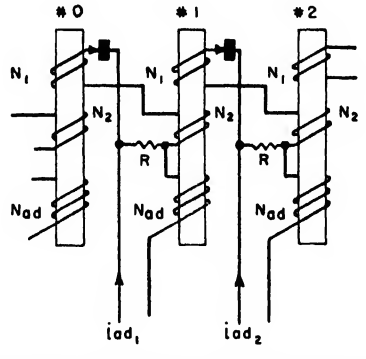


FIG. 10. Another form of delay line of two cores per digit.

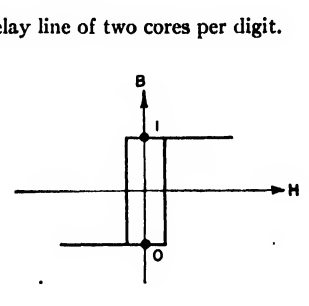


FIG. 11. Ideal hysteresis loop.

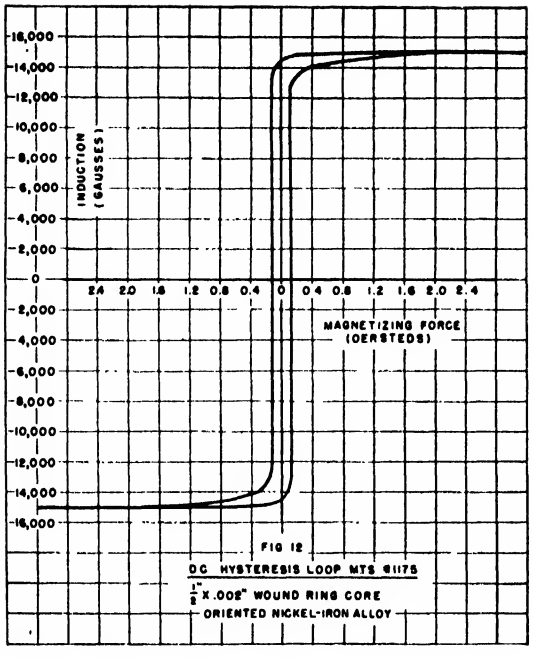


FIG. 12. Typical hysteresis loop of materials used.

3. *Counter.* Since the information stored in the delay line must be stepped along by advancing pulses, it is possible to register a single digit at the beginning of the line, and by the location of this digit along the line at a later instant to determine the number of advancing pulses that have been applied. So far the speed of counting is limited to about 50 kc. This, however, is sufficiently rapid to find wide applications.

4. *Telegraphy.* Because the speed of propagation is variable, a magnetic delay line may serve as the terminal storage for a high speed telegraphy system. The information on a high speed telegraphic line can be distributed to several delay lines. While some lines are being filled, any one of them can be discharged much more slowly to a separate line or directly to a typewriter. Similarly, several slow speed information sources can be read into the magnetic delay lines and then be fed into a high speed telegraphic line by electronic switching control of the advancing pulses.

VIII. ACKNOWLEDGMENTS

The magnetic core materials employed in this research were supplied by the Allegheny Ludlum Steel Corporation, whose cooperation and assistance the authors wish to acknowledge. The use of magnetic cores as a means of obtaining static magnetic storage was originally proposed by Howard H. Aiken, director of the Computation Laboratory, who has worked closely with the authors.

A General Solution for the Bending of Beams on an Elastic Foundation of Arbitrary Continuity*

M. Hetényi
Northwestern Technological Institute, Evanston, Illinois
(Received July 29, 1949)

In the analysis of beams which are supported elastically along their entire length, it is usually assumed that the elastic support is provided either by a series of independent springs, or by a completely continuous elastic medium.

The present paper proposes a method of solution for the intermediate cases, that is, for foundations of partial continuity, by assuming that such foundations may be represented by beams imbedded into an otherwise discontinuous supporting material. In a foundation of this type, there are three independent elastic constants, the resilience of the discontinuous spring layers above and below the beam that is within the foundation, and the flexural rigidity of the foundation beam itself. By changing the values of these three constants, elastic foundations of a wide variety may be produced or, conversely, the constants may be adjusted in such a manner as to represent approximately the characteristics of any given elastic supporting medium.

In the paper a general procedure is presented from which solutions can be derived for the bending of beams on such foundations under any loading or end conditions. The method is illustrated by examples showing the range of pressure distributions that may occur under beams supported on foundations of this type.

IN dealing with problems of bending of beams which are supported elastically along their entire length, it is usually assumed that the supporting medium is provided either by a series of independent springs, or by a completely continuous elastic body. Solutions for a large variety of problems were developed in the past on the basis of either of these two assumptions.† There remains an obvious need, however, for a method of analyzing the intermediate cases, where the foundation material may exhibit an arbitrary and partial degree of continuity. The aim of the present paper is to propose a method of solution for problems of this kind.

A partially continuous foundation will be assumed in this analysis as composed of a bending resistant beam, imbedded between two layers of otherwise independent elastic springs. A foundation of this type can be characterized by three elastic constants, the k_1 and k_2 spring constants (lb./in. per inch) of the two layers, and the flexural rigidity EI_2 of the beam within the foundation. If a trial load q lb./in. is placed along a length b on the surface of such a composite foundation, the resulting deflection line may also be defined by three data, namely, by the amount of local deflection δ_1 within the loaded region, by the maximum deflection ordinate δ_2 of the continuous part of the displacement curve, and by the distance d at which displacements vanish [Fig. 1(a)]. The three characteristic constants of the foundation, k_1, k_2, and EI_2, can be calculated from the observation data, δ_1, δ_2, and d, on the basis of the following relationships in the known solution of a single beam on an elastic foundation without continuity: $\delta_1=q/k_1$, $\delta_2=qb(\lambda_2/2k_2)$, and $d=3\pi/4\lambda_2$, where $\lambda_2=(k_2/4EI_2)^{\frac{1}{4}}$.

If a load carrying beam of flexural rigidity EI_1 is placed on a foundation defined by these three constants, k_1, k_2, and EI_2, as shown in Fig. 1(b), the differential equations of bending may be derived as follows: denoting the deflection ordinates of the upper and the lower beam by y_1 and y_2, respectively, we find that the pressure in the foundation under the upper beam is $p_1=k_1(y_1-y_2)$, and under the lower beam $p_2=k_2y_2$, while the resultant pressure acting on the lower beam is $p_2-p_1=(k_1+k_2)y_2-k_1y_1$. Since these pressures are continuously distributed along the two beams, we have from the flexural theory:

$$EI_1(d^4y_1/dx^4)=-p_1=-k_1(y_1-y_2) \qquad \text{(a)}$$

and

$$EI_2(d^4y_2/dx^4)=-(p_2-p_1)=-(k_1+k_2)y_2+k_1y_1. \qquad \text{(b)}$$

The pressures in these cases are considered positive when accompanied by positive (downward) deflections. From (a) we have

$$y_2=\frac{EI_1}{k_1}\frac{d^4y_1}{dx^4}+y_1 \qquad \text{(c)}$$

or

$$\frac{d^4y_2}{dx^4}=\frac{EI_1}{k_1}\frac{d^8y_1}{dx^8}+\frac{d^4y_1}{dx^4}. \qquad \text{(d)}$$

On substituting these expressions into (b), y_2 will cancel out, and we have the differential equation of the elastic line of the upper beam

$$(d^8y_1/dx^8)+2\alpha(d^4y_1/dx^4)+(\alpha_2-\beta_2)y_1=0,$$

where

$$\alpha=[k_1(I_1+I_2)+k_2I_1]/2EI_1I_2$$

and

$$\beta=[\alpha^2-(k_1/EI_1)(k_2/EI_2)]^{\frac{1}{4}}.$$

* Paper presented at the VII International Congress of Applied Mechanics, London, England, September 5–11, 1948.

† See section on references at the end of this paper.

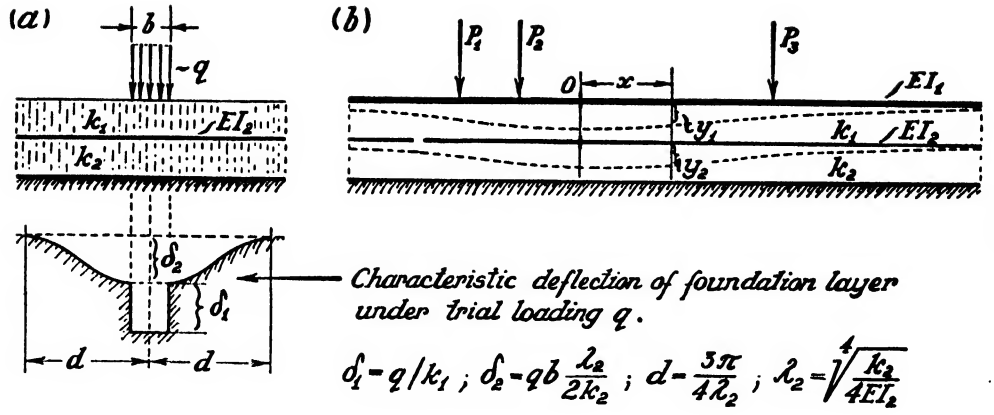


FIG. 1. Foundation layer with an arbitrary degree of continuity.

The general solution of (e) can be written in the form:

$$y_1 = e^{\lambda_1 x}(C_1\cos\lambda_1 x + C_2\sin\lambda_1 x) + e^{-\lambda_1 x}(C_3\cos\lambda_1 x + C_4\sin\lambda_1 x) + e^{\lambda_2 x}(C_5\cos\lambda_2 x + C_6\sin\lambda_2 x) + e^{-\lambda_2 x}(C_7\cos\lambda_2 x + C_8\sin\lambda_2 x), \quad (1)$$

where

$$\lambda_1 = [(\alpha+\beta)/4]^{\frac{1}{4}} \quad \text{and} \quad \lambda_2 = [(\alpha-\beta)/4]^{\frac{1}{4}}.$$

Substituting the solution from (1) into (c), we have the expression for the elastic line of the beam within the foundation:

$$y_2 = [1-(\alpha+\beta)(EI_1/k_1)][e^{\lambda_1 x}(C_1\cos\lambda_1 x + C_2\sin\lambda_1 x) + e^{-\lambda_1 x}(C_3\cos\lambda_1 + C_4\sin\lambda_1 x)] + [1-(\alpha-\beta)(EI_1/k_1)] \times [e^{\lambda_2 x}(C_5\cos\lambda_2 x + C_6\sin\lambda_2 x) + e^{-\lambda_2 x}(C_7\cos\lambda_2 x + C_8\sin\lambda_2 x)]. \quad (2)$$

Expressions for the slope $\theta = dy/dx$, bending moment $M = -EI(d^2y/dx^2)$, and shearing force $Q = -EI(d^3y/dx^3)$ may be obtained for each of the above beams by differentiating the corresponding equation of the elastic line, (1) or (2), respectively.

In applying the above general solutions to the analysis of beams of finite length, the scheme illustrated in Fig. 2 may be the most conveniently used. Here it is first assumed that the beam within the foundation is severed at A and B, resulting in relative displacements and rotations between the severed ends, as the upper beam is loaded. The continuity of slope and deflection in the lower beam is then re-established by the application of end forces P_0 and moments M_0 in the manner shown in the figure.

In using this method of calculation, one needs to know the particular solutions of the elastic lines corresponding to each of the loading cases shown in Fig. 2. The integration constants, C_1 to C_8, in the general solutions (1) and (2) were determined from the prevailing end conditions in each of these particular cases of loading, giving the following results.

Concentrated load P, centrally applied at the upper beam (Fig. 3).

$$C_1 = A\left[\frac{\cosh^2\lambda_1 c + \cos^2\lambda_1 c}{\sinh 2\lambda_1 c + \sin 2\lambda_1 c} - \frac{1}{2}\right]$$

$$C_2 = A\left[\frac{1}{2} - \frac{\sinh^2\lambda_1 c + \sin^2\lambda_1 c}{\sinh 2\lambda_1 c + \sin 2\lambda_1 c}\right]$$

$$C_3 = A\left[\frac{\cosh^2\lambda_1 c + \cos^2\lambda_1 c}{\sinh 2\lambda_1 c + \sin 2\lambda_1 c} + \frac{1}{2}\right]$$

$$C_4 = A\left[\frac{1}{2} + \frac{\sinh^2\lambda_1 c + \sin^2\lambda_1 c}{\sinh 2\lambda_1 c + \sin 2\lambda_1 c}\right]$$

$$C_5 = -B\left[\frac{\cosh^2\lambda_2 c + \cos^2\lambda_2 c}{\sinh 2\lambda_2 c + \sin 2\lambda_2 c} - \frac{1}{2}\right]$$

$$C_6 = -B\left[\frac{1}{2} - \frac{\sinh^2\lambda_2 c + \sin^2\lambda_2 c}{\sinh 2\lambda_2 c + \sin 2\lambda_2 c}\right]$$

$$C_7 = -B\left[\frac{\cosh^2\lambda_2 c + \cos^2\lambda_2 c}{\sinh 2\lambda_2 c + \sin 2\lambda_2 c} + \frac{1}{2}\right]$$

$$C_8 = -B\left[\frac{1}{2} + \frac{\sinh^2\lambda_2 c + \sin^2\lambda_2 c}{\sinh 2\lambda_2 c + \sin 2\lambda_2 c}\right]$$

where:

$$A = \left[\frac{k_1}{EI_1} - \alpha + \beta\right]\frac{P}{EI_1}\,\frac{1}{16\beta}\,\frac{1}{\lambda_1^3}$$

and

$$B = \left[\frac{k_1}{EI_1} - \alpha - \beta\right]\frac{P}{EI_1}\,\frac{1}{16\beta}\,\frac{1}{\lambda_2^3}.$$

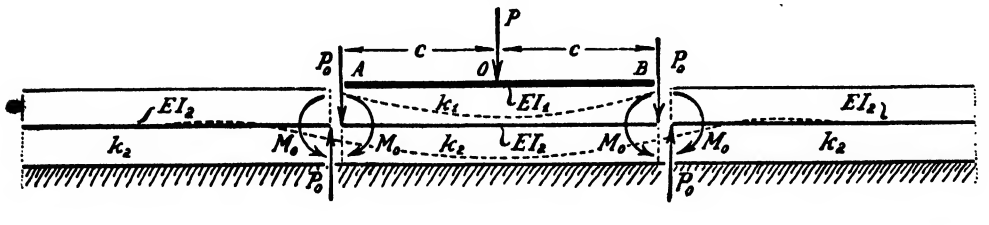


FIG. 2. Method of analyzing a beam of finite length A–B, supported on a foundation layer.

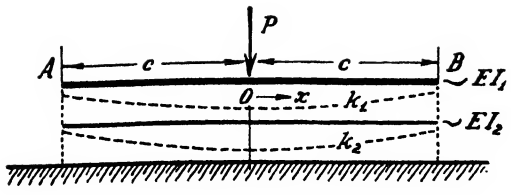


FIG. 3. Particular solution to problem shown in Fig. 2.

Lower beam subjected to equal concentrated end forces P_0 (Fig. 4).

$$C_1=C_3=-\frac{P_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_1^3}\frac{1}{}\frac{\cosh\lambda_1 c\cos\lambda_1 c}{\sinh 2\lambda_1 c+\sin 2\lambda_1 c}$$

$$C_2=-C_4=-\frac{P_0}{EI_1}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_1^3}\frac{1}{}\frac{\sinh\lambda_1 c\sin\lambda_1 c}{\sinh 2\lambda_1 c+\sin 2\lambda_1 c}$$

$$C_5=C_7=\frac{P_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_2^3}\frac{1}{}\frac{\cosh\lambda_2 c\cos\lambda_2 c}{\sinh 2\lambda_2 c+\sin 2\lambda_2 c}$$

$$C_6=-C_8=\frac{P_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_2^3}\frac{1}{}\frac{\sinh\lambda_2 c\sin\lambda_2 c}{\sinh 2\lambda_2 c+\sin 2\lambda_2 c}.$$

Lower beam subjected to equal concentrated end moments M_0 (Fig. 5).

$$C_1=C_3=\frac{M_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_1^2}\frac{1}{}\frac{\sinh\lambda_1 c\cos\lambda_1 c-\cosh\lambda_1 c\sin\lambda_1 c}{\sinh 2\lambda_1 c+\sin 2\lambda_1 c}$$

$$C_2=-C_4=\frac{M_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_1^2}\frac{1}{}\frac{\sinh\lambda_1 c\cos\lambda_1 c+\cosh\lambda_1 c\sin\lambda_1 c}{\sinh 2\lambda_1 c+\sin 2\lambda_1 c}$$

$$C_5=C_7=-\frac{M_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_2^2}\frac{1}{}\frac{\sinh\lambda_2 c\cos\lambda_2 c-\cosh\lambda_2 c\sin\lambda_2 c}{\sinh 2\lambda_2 c+\sin 2\lambda_2 c}$$

$$C_6=-C_8$$

$$=-\frac{M_0}{EI_2}\frac{k_1}{EI_1}\frac{1}{4\beta\lambda_2^2}\frac{1}{}\frac{\sinh\lambda_2 c\cos\lambda_2 c+\cosh\lambda_2 c\sin\lambda_2 c}{\sinh 2\lambda_2 c+\sin 2\lambda_2 c}.$$

Semi-infinite beam on a simple elastic foundation loaded by concentrated end force P_0 and moment M_0 (Fig. 6). Equation of elastic line:

$$y=\frac{2P_0\lambda}{k_2}e^{-\lambda x}\cos\lambda x-\frac{2M_0\lambda^2}{k_2}e^{-\lambda x}(\cos\lambda x-\sin\lambda x),$$

where $\lambda=[(k_2/4EI_2)]^{\frac{1}{4}}$.

The wide range in the pressure distribution types which may be covered by this theory is illustrated in Fig. 7, where a centrally loaded beam of length $2c=280$ in. and flexural rigidity $EI=250\times10^6$ lb. in.^{2} is placed on foundations with increasing degree of continuity. On the left side of the figure the characteristic deflection curves are shown for each of the three types of foundation, as they are subjected to a trial load of $q=10$ lb./in. distributed along a length of 10 in.

The first type of foundation lacks continuity entirely, and its spring constant can be calculated through dividing the applied trial load q by the observed local deflection ordinate $k=10/0.75=13.33$ lb./in.^{2}. The second type of foundation is characterized by a local deflection of $\delta=0.25$ in. and a continuous deflection curve of maximum ordinate of $\delta=0.50$ in. and wave-length $d=50$ in., which define the following elastic constants for the foundation layer: $k_1=20$ lb./in.^{2}, $k_2=8.44$ lb./in.^{2}, and $EI_2=0.476\times10^8$ lb. in.^{2}. The third type of foundation is assumed to have the greatest degree of continuity, with two-thirds of the maximum deflection being of the continuous type, $\delta_1=0.25$ in. and $\delta_2=0.50$ in., and the effect of continuity being observable up to a distance of $d=100$ in. from the center of application of the trial load, which define the following elastic constants for the foundation layer: $k_1=40$ lb./in.^{2}, $k_2=2.36$ lb./in.^{2}, and $EI_2=1.90\times10^6$ lb. in.^{2}.

The pressure distribution under the loaded beam for each of the average pressure $p_0=P/2c$. It is to be noted that the flexural values of the loaded beam ($2c=280$ in., $EI_1=250\times10^6$ lb. in.^{2}) and the total compressibility of the foundations ($\delta_1+\delta_2=0.75$ in.) were the same in each case, thus the pronounced difference in the resulting pressure distributions was due solely to the different degrees of continuity in the material of the foundation.

The outlined method of analysis may be applied to any elastic supporting medium (subsoils, elastic layers, packing or gasket materials etc.) which exhibits a partly local and partly continuous deflection curve when tested under localized pressure. Though in the paper

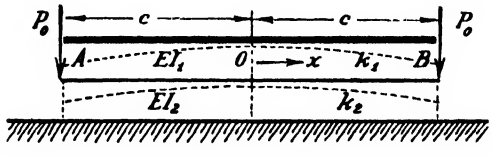


FIG. 4. Particular solution to problem shown in Fig. 2.

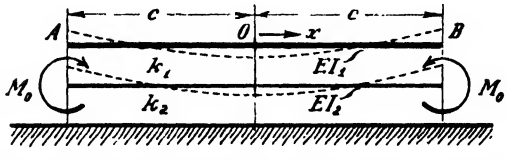


FIG. 5. Particular solution to problem shown in Fig. 2.

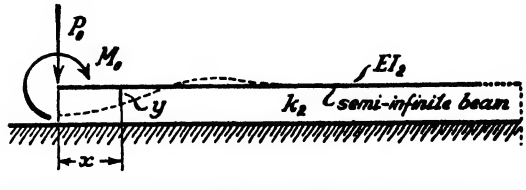


FIG. 6. Particular solution to problem shown in Fig. 2.

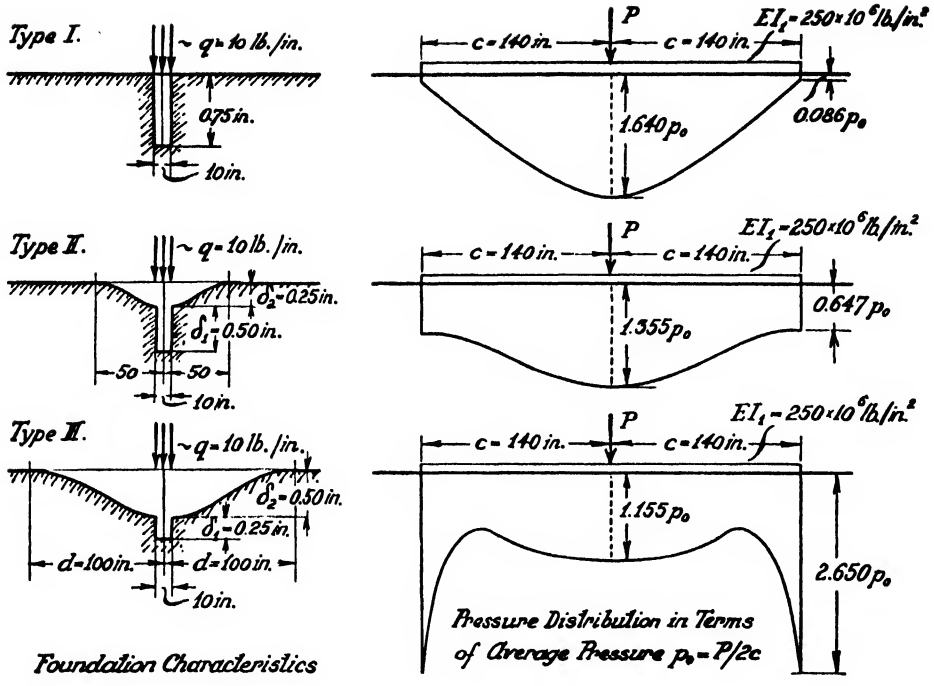


FIG. 7. Pressure distribution under a centrally loaded beam, supported on foundations with various degrees of continuity.

only one-dimensional problems were considered, when beam and foundation are of the same constant width, the same principle may be readily extended to the analysis of corresponding two-dimensional plate problems.

REFERENCES

The discontinuous type of elastic foundation was first introduced by E. Winkler [*Die Lehre von der Elastizität und Festigkeit* (Prague, 1867)]. This type of foundation was assumed in the rail investigations by H. Zimmermann [*Die Berechnung des Eisenbahnoberbaues* (Berlin, 1888)], and a large variety of solutions developed on this basis were discussed by K. Hayashi [*Theorie des Trägers auf elastischer Unterlage* (Berlin, 1921)].

Problems of beams supported on an elastic continuum were investigated, among others, by K. Wieghardt [Zeits. f. angew. Math. Mech. 3, 165–184 (1922)], M. A. Biot [J. App. Mech. A, 1–7 (1937)], and G. Bosson [Phil. Mag. 37–50 (1939)].

A comprehensive study of both types of foundations and the first exposition of the concept of partially continuous foundations may be found in the author's book *Beams on Elastic Foundation* (University of Michigan Press, Ann Arbor, 1946).

Letters to the Editor

Microwave Frequency Dividers

HAROLD LYONS
National Bureau of Standards, Washington, D. C.
August 8, 1949

IN order to extend the range of application of microwave techniques it is becoming more and more important to be able to carry out the operation of frequency division in the microwave region just as at lower frequencies. Such needs are felt, particularly, in connection with spectroscopic or atomic clocks and frequency standards which usually use spectrum lines in the neighborhood of 3000–20,000 Mc or higher.[1] It is therefore necessary to look for a frequency division principle which uses components and techniques that are practicable in the microwave range.[2]

One could inject a signal to be divided into an oscillator operating at a sub-multiple of the injected frequency and lock it in.[3] This well-known principle has obvious disadvantages. If control is lost the lower frequency oscillator will run free. A frequency divider which avoids this difficulty and which is widely used at low frequencies, e.g., in precision quartz-crystal clocks, is the regenerative modulator type.[4] This circuit also works well even in the presence of high noise levels and distorted input signals. It can be designed so that the output will be linearly dependent on the input and maintain a constant phase angle with respect to it. It can generate both sub-multiple and other fractional frequency ratios. Such a circuit to divide down from 9300 Mc to 3100 Mc has been built and tested.[5] A block diagram is shown in Fig. 1. Here it is seen that the 9300 Mc signal is applied to a mixer; to the mixer is also applied a signal at 6200 Mc. This may be a noise component or transient generated by turning the circuit on. The 6200 Mc signal beats with the 9300 Mc input signal to give an output of 3100 Mc. This is amplified by means of the Sperry 2K35 klystron amplifier tubes, and then fed to a Sperry 2K46 klystron multiplier which is tuned to double the input signal to 6200 Mc. After multiplication the signal is amplified at 6200 Mc by means of a Sperry SAC-19 klystron amplifier and fed back to the mixer. If the gain around the feedback loop is great enough, the output signal at 3100 Mc will be maintained independently of the phase shift in the feedback loop. No output will appear unless there is an input signal since the 3100 Mc beat note can only be obtained in this way. If the gain is not high enough in the regenerative circuit it may be necessary to start the operation by injecting some sort of transient. In the circuit shown this was not necessary. Since no attempt was made in this first circuit to design a very efficient mixer, it may not be necessary to use all of the amplification shown in Fig. 1. In principle, the 2K46 frequency multiplier could be replaced by any nonlinear modulator device, such as a crystal rectifier if enough gain were available and the signal level high enough.

Among numerous tests conducted, the input, intermediate, and output frequencies were measured and found to track each other and it was also shown by the insertion of wave-guide filters that no other output signals were present except those in the 3000 Mc range. The output signal was taken from the middle cavity of one of the 2K35 tubes. A definitive test of the accuracy of frequency division was made as follows. The 9300 Mc input signal was fed to a spectrum analyzer where it was viewed on the scope as a pip. The output signal was monitored by means of a frequency meter and second spectrum analyzer. In addition it was fed to a crystal multiplier which multiplied by three and then fed to the same spectrum analyzer as the input frequency. If the output frequency times three is exactly equal to the input frequency, the two pips on the scope should coincide. The two individual pips could be identified by first displaying one and then the other by control of their intensities by means of attenuators. When the two pips were superimposed they blended smoothly into one larger pip; this was done under conditions of highest dispersion in the spectrum analyzer. No beating of the two signals whatsoever could be discerned nor any shift or change in shape of the pulses. If it is assumed rather arbitrarily that a beat note of approximately one cycle per second could be observed, this means that the output frequency times three was equal to the input frequency with an accuracy of at least one part in 10^{10}. In this type of test no difficulty results if the input-oscillator-frequency slowly drifts back and forth since the output frequency drifts with it. It is clear that direct frequency calibrations of extreme accuracy would be necessary to obtain a test with the sensitivity of the one just described.

Since another klystron multiplier tube, the Sperry 2K47, is commercially available and multiplies from approximately 270 Mc up to the 3000 Mc range, frequency dividers could be built to go down to as low a frequency as desired. It is important to note that in this type of circuit it is not necessary to have amplifiers available at the input frequency but only at a lower frequency, essentially a divided frequency. Since klystron multipliers are very efficient and much easier to make than amplifiers, a multiplier tube and frequency divider dividing down from the 24,000 Mc region, rich in spectrum lines, should be possible.

Since amplifiers in this type of circuit are only needed at the divided frequency, some interesting possibilities are opened up by consideration of circuits of the type shown in Fig. 2. In this circuit we make the regenerative-modulator divider provide its own input by providing a second regenerative channel. In this case, if there is enough gain, the circuit should oscillate and divide in frequency at the same time. A more general arrangement allowing more flexibility in the choice of frequencies which can be used with any given set of tubes is shown in Fig. 3. Figures 2 and 3 show a proposed application to an atomic oscillator and atomic clock.[6] Here by atomic oscillator we mean a self-excited or feedback oscillator, the frequency of which is directly controlled by an absorption cell. This is in contra-distinction to the stabilized servo mechanism or frequency discriminator type of spectrum-line controlled oscillators.[1,6] To make an atomic clock it is then necessary to divide down by means of straight frequency division from the atomic oscillator to a frequency low enough to drive the usual synchronous motor type of clock. Two types of atomic oscillators in which the

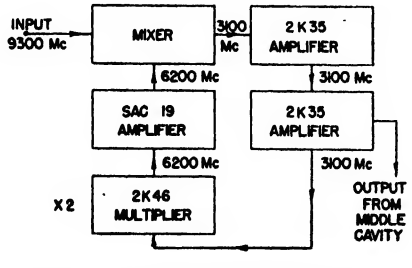


FIG. 1. Microwave regenerative modulator frequency divider, dividing 3 to 1.

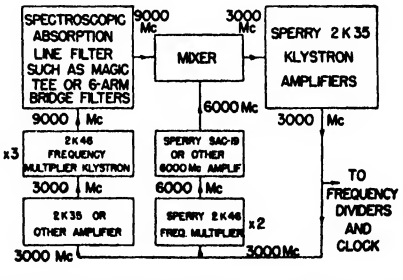


FIG. 2. Proposed combination atomic oscillator and frequency divider, using amplifiers at divided frequencies, applied to an atomic clock.

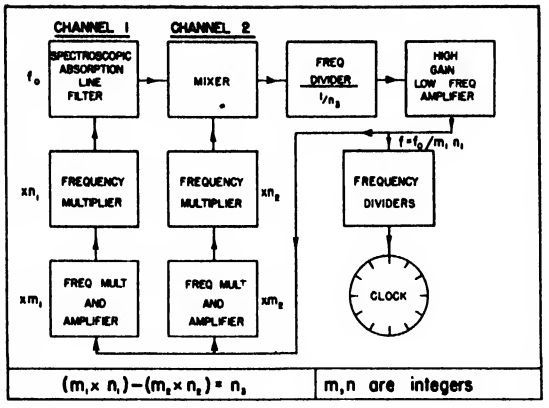


FIG. 3. General and more flexible block diagram of atomic clock using combination oscillator and divider.

feedback frequency is controlled by means of spectroscopic circuits have been described.[1,2,5] Here the spectroscopic filter is based on the use of a magic-tee or a six-arm, wave-guide Wheatstone bridge in which an absorption cell containing ammonia or other absorbing gas at low pressure determines the frequency which will be transmitted by the filter. As shown in Fig. 3, this spectroscopic filter transmits an input frequency to the regenerative modulator only at the frequency of the spectrum line used in the filter. This, therefore, controls the frequency of the oscillator and similarly that of the divided frequency. However, as compared to the atomic oscillators previously described, in this method not only is frequency division carried on simultaneously with controlled oscillation, but an amplifier at the spectrum line frequency is no longer needed.

Although it is very desirable to go back up to the 3,3 line of ammonia at approximately 24,000 Mc, and we are attempting development and procurement of suitable multiplier and amplifier tubes for this purpose, an atomic clock could be developed using a spectrum line in the 9000 Mc range using available amplifier and multiplier tubes and the circuit shown in Fig. 2. For this purpose the various deuterated species of ammonia were prepared and spectrum lines searched for in order to provide a strong absorption line at a lower frequency. Although this work is still incomplete, about 30 lines have been found in the range between 3000 and 10,000 Mc.[5] Such spectrum lines can be used to make a clock following the circuit shown in Fig. 2 or simply used in a separate atomic oscillator driving a frequency divider.

[1] "The Atomic Clock, an Atomic Standard of Frequency and Time," NBS Technical News Bulletin 33, (February, 1949). Also, Harold Lyons, Phys. Rev. 74, 1203 (1948).
[2] "Microwave spectroscopic frequency and time standards," Harold Lyons, presented at the AIEE-IRE-NBS Conference on High Frequency Measurements, Washington, D. C., January, 1949. Also, Electrical Engineering 68, 251 (1949).
[3] E. Norrman, Proc. I.R.E. 34, 799 (1946).
[4] R. L. Miller, Proc. I.R.E. 27, 446 (1939). Also, W. A. Morrison, Bur. Stand. Tech. J. 27, 510 (1948).
[5] Harold Lyons, Phys. Rev. 76, 161 (1949).
[6] a. Chaffee, Fletcher and Cooke, Harvard University Progress Reports on Contract N5-ORI-76, 1947–1949. b. R. V. Pound, Rev. Sci. Inst. 17, 490 (1946); Smith, DeQuevedo, Carter, and Bennett, J. App. Phys. 18, 1112 (1947); DeQuevedo and Smith, J. App. Phys. 19, 831 (1948); c. W. D. Hershberger and L. E. Norton, RCA Rev. 9, 38 (1948). d. C. H. Townes, A. N. Holden, and F. R. Merritt, Phys. Rev. 74, 1113 (1948).

Particle Size of Evaporated Gold

P. G. WILKINSON AND L. S. BIRKS
U. S. Naval Research Laboratory, Washington, D. C.
October 5, 1949

MANY observers have noted that the gold on shadowed electron micrographs granulates when exposed to an intense beam of electrons. It has been generally assumed that before electron bombardment, the gold was a smooth continuous film. Picard and Duffendack[1] published micrographs of gold evaporated at 10^{-6} mm pressure in which the gold appeared as discreet particles. From work[2] we have done on correlating electrical resistance properties of gold with microstructure, we believe that such is generally the condition of evaporated gold except when contaminated by the presence of tungstic oxide. In the usual bell jar conditions for gold shadowing electron micrographs at about 10^{-3} mm pressure, the deposit contained about five percent tungsten by analysis. This gave a characteristic matte background which was observed no matter how careful the operator was to use low beam intensity and fast exposure. When the percentage of tungstic oxide was reduced either by pumping to lower pressure or by flushing the system during evaporation with a few microns of purified nitrogen, free of oxygen, the gold appeared as discreet particles, similar to those observed by Picard and Duffendack. Such deposits showed no tungsten by chemical analysis. Figure 1

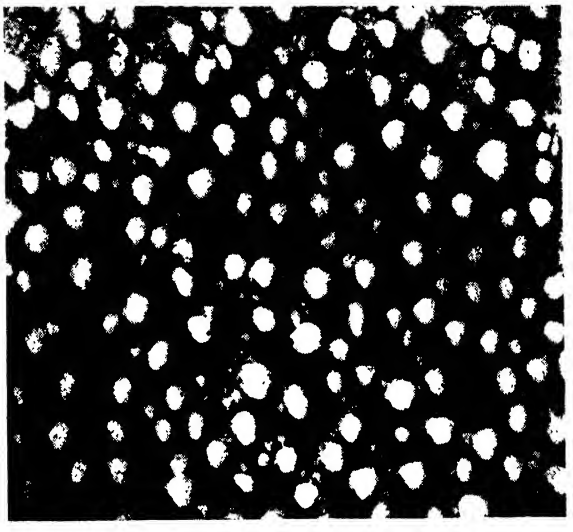


FIG. 1. Gold evaporated at 10^{-6} mm pressure and shadowed with tungstic oxide.

proves the existence of discreet gold particles before electron bombardment. The gold was evaporated at 10^{-6} mm pressure and then shadowed with tungstic oxide before it was placed in the microscope. The shadows could not have appeared if the granulation had taken place in the electron microscope.

Figure 1 is typical of all the tungsten-free gold deposits we have observed.

[1] R. G. Picard and O. S. Duffendack, J. App. Phys. 14, 291 (1943).
[2] P. G. Wilkinson and L. S. Birks, J. App. Phys. (to be published); P. G. Wilkinson and L. S. Birks, N.R.L. Report No. P-3443.

Energy Partition in Isothermal Fracture*

MARGUERITE M. ROGERS**
The University of North Carolina, Chapel Hill, North Carolina
AND
F. T. ROGERS, JR.†
Oak Ridge National Laboratory, Oak Ridge, Tennessee
October 14, 1949

MEASUREMENTS have been made of the elastic potential energy (e) and absorbed energy (w) associated with the fracture of four-notch specimens (0.70″×0.75″×4.6″ with 15 percent Izod notches) of a special aluminum-killed mild steel (Bethelehem), by observing for several values of deformation (s),

TABLE I. Relative properties of pre-brittle and ductile specimens, expressed as ratios derivatives for particular pre-brittle specimens to derivatives for the mean of ductile specimens.

x (in.)	Specimen number, B-series	dx_B/dx_D	dw_B/dw_D	G_B/G_D
0.02	208	1.2	0.8	1.1
0.04		1.6	0.3_8	1.0
0.058*		9	0.1_8	0.4
0.04	203	1.10	0.76	1.0_8
0.058		1.13	0.81	1.0_8
0.10		1.16	0.93	1.0_6
0 14		1.19	0.95	0.9_8
0.175		1.54	0.83	0.9_4
0.22		1.90	0.56	0.8_2
0.245		9.5	0.21	0.5
0.04	237	1.10	0.71	1.0_8
0.058		1.13	0.75	1.0_8
0.10		1.16	0.78	0.9_8
0.14		1.54	0.72	0.9_4
0.175*		8.1	0.20	0.7

* Brittle fracture occurred 0.005 in. in excess of these values, for the respective specimens.

the fracture depth (x), and the applied dynamic load. Specimens were held thermostatically at temperatures between 0°F and 50°F so that strain ageing would be small, and deformation was effected very slowly at a constant rate. The essential features of the apparatus have been described elsewhere.[1] At the higher temperatures the specimens fractured in a controlled (ductile) manner, but at the lower ones many ultimately underwent uncontrolled (brittle) fracture during test; however, temperature did not affect the data measureably.

When plotted against s, the data for ductile and pre-brittle specimens were identical. But when plotted against x, the data clearly identified the pre-brittle specimens as different from the ductile ones which were chemically and metallurgically identical with them. As Table I shows (in terms of derivatives), brittle fracture was preceded in the pre-brittle phases by excessive fracture depths and by deficient (in terms of x) energy absorption (see columns 3 and 4, respectively). The three brittle specimens reported in Table I are representative of the many others observed.

Column 5 in Table I refers to the function

$$G=1+de/dw. \quad (1)$$

G. R. Irwin has suggested[2] that a controlled fracture should become uncontrolled when $G\rightarrow 0$, by extending the concepts of A. A. Griffith.[3] Although the present data are not sufficiently precise to provide an ultimate test of Irwin's suggestion, they are in entire accord with it as far as they go, and to that extent serve to emphasize the significance of energy partition in macrofracture phenomena.

* The experimental portions of this work were carried out in the research laboratories of The University of North Carolina under the provisions of Contract No. N6-ori-227 (and Amendment I) of the ONR with the University. Dr. P. E. Shearin and Mr. R. M. Trimble provided material assistance.

** Address at the time of writing: Oak Ridge National Laboratory, Oak Ridge, Tennessee.

† Present address: U. S. Naval Ordnance Test Station, Inyokern, California.

[1] Shearin, Ruark, and Trimble, *Strength of Solids* (Physical Society, London, 1947), p. 158; *Fracturing of Metals* (Am. Soc. Metall., 1948), p. 167.
[2] G. R. Irwin, *Fracturing of Metals* (Am. Soc. Metall., 1948), p. 147.
[3] A. A. Griffith, Phil. Trans. Roy. Soc. (London) 221, 163 (1920).

A Low Temperature Replica Method for Electron Microscopy

C. E. HALL
Massachusetts Institute of Technology, Cambridge, Massachusetts
September 6, 1949

A REPLICA method has been tested which appears to be of value for the electron microscopy of aqueous substances while they are in the frozen state. The apparatus employed is shown diagrammatically in Fig. 1. The specimen, A, on glass is supported in a vacuum chamber in contact with a copper block, B, which is pre-cooled to liquid air temperature. The block is suspended from a glass rod and can be set at any desired angle by means of a conical vacuum plug. When a vacuum has been obtained, the temperature of the specimen is temporarily raised to about −60° to −90°C with radiant heat from an external microscope lamp in order to sublime ice crystals which may have condensed on the specimen from the atmosphere and also to secure a surface etch. In the absence of an intense source of radiant heat, the conductivity of the supporting glass is sufficient to maintain the specimen at a temperature close to that of the cold sink. After the etching procedure, the specimen is shadowed with Cr and a supporting film of Si evaporated from a BeO crucible[1] is deposited on the surface at normal incidence. (Although Si was used on these experiments, SiO has subsequently proved more convenient for evaporated films since it does not attack tungsten.) Finally, the cold block is removed from the vacuum, the glass plate is detached and the surface film is floated on water after the specimen has warmed to room temperature.

In order to test the feasibility of the method a dilute aqueous suspension of silver halide particles from a photographic plate was used as a test object. Silver halide was removed by floating the replica film on sodium hyposulfite. A micrograph of such a film is shown in Fig. 2. Although the etch was too light to cause the large particles to project far above the surface their outline may be discerned. The surrounding surface of frozen, dilute gelatin displays a crater-like structure, the significance of which is not understood at present. No evidence of gross ice crystals has been observed within quick-frozen specimens.

The micrograph in Fig. 3, for which the etching procedure was omitted, shows ice crystals condensed from air on the surface of frozen distilled water. The micrograph demonstrates the feasibility of obtaining replicas of ice surfaces in this manner and also emphasizes the necessity for either carefully shielding the cold specimen surface or removing the condensed moisture by sublimation before the process is employed.

Since the method depends on the immobilization of a water surface in a vacuum, it is of interest to estimate the rate at which water will sublime from a specimen. The rate of sublimation can be computed[2] for a given temperature from a knowledge of the equilibrium vapor pressure of ice which is tabulated[3] to −98°C and extrapolated values may be obtained with the aid of Clapey-

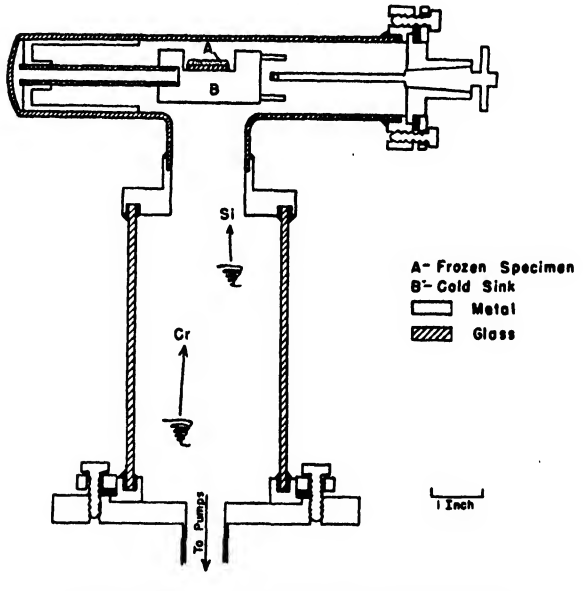


FIG. 1. Apparatus for low temperature replica method.

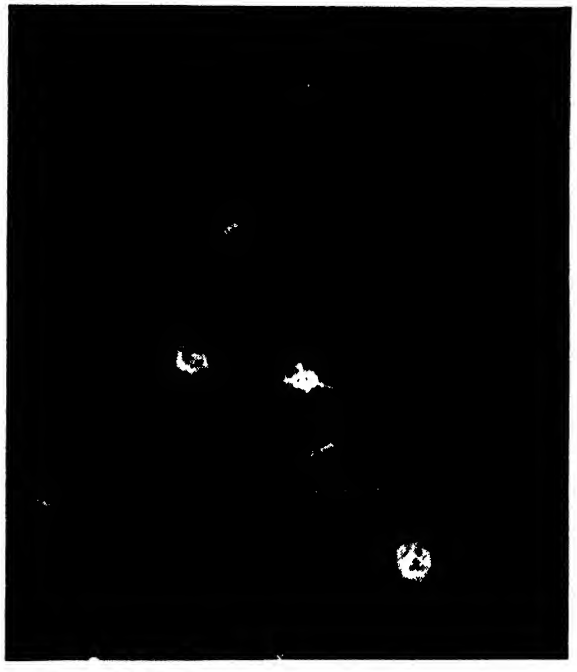


FIG. 2. Low temperature replica of surface of a frozen dilute suspension of silver halide particles.

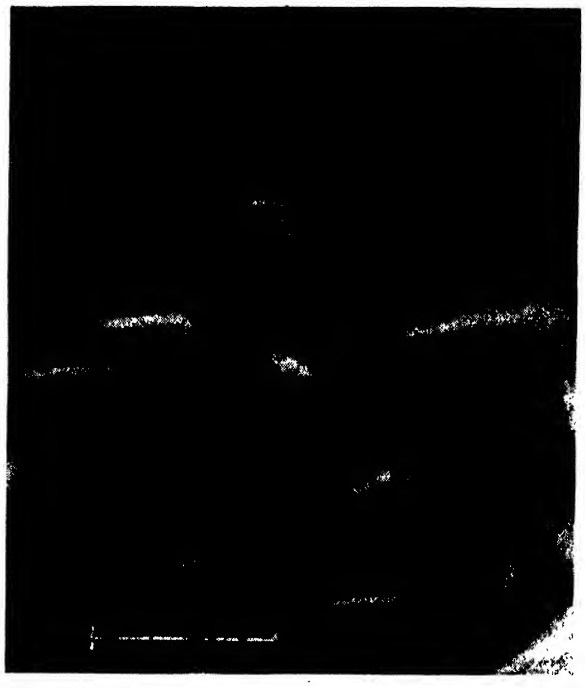


FIG. 3. Low temperature replica of ice crystal condensed on frozen distilled water surface, shadow-cast with Cr, but not etched.

ron's equation.[4] The vapor pressures together with the computed rates at which an ice surface would recede in a perfect vacuum at various temperatures of interest are listed in Table I. It is apparent from the table that if the surface is to be maintained over periods of several minutes, the temperature should be in the range −120°C or lower, and that for the etching procedure, the rate of sublimation is reasonably rapid at about −90°C and higher.

TABLE I.* Estimated rate of recession of an ice surface in a vacuum.

Temperature °C	Vapor pressure dynes cm^{-2}	Rate of recession A sec.$^{-1}$
−90	9.3×10^{-3}	140
−100	1.4×10^{-3}	23
−110	1.6×10^{-4}	2.6
−120	1.3×10^{-5}	0.2
−130	7.7×10^{-7}	0.01

* Vapor pressure at −90°C from *Handbook of Chemistry and Physics*; other values calculated from Clapeyron's equation assuming latent heat of sublimation 675 calories per g.

With reasonable assumptions concerning the conductivity between specimen and cold sink, estimates can be made of the rise in specimen temperature due to radiant heat from the hot evaporation filaments and due to heat released through deposition of evaporated shadowing and film materials. Such estimates indicate that under typical conditions the temperature rises about 2°C from radiant heat and is also of negligible magnitude due to condensing Si or shadowing metal. Heat is also absorbed during transfer from liquid air to the vacuum chamber by condensation of moisture from the air and conduction to the atmosphere. When the operation was performed with reasonable speed, the temperature of the block (measured with a platinum resistance thermometer) had usually risen to about −145°C by the time the pressure in the vacuum chamber had reached the point where the oil diffusion pump became effective. The temperature rose at the rate of about 1°C per minute beyond this point. Clearly, in order to complete the process before the specimen warms to a temperature where an ice surface becomes unstable, it is necessary to transfer the cold block rapidly to the vacuum chamber and attain a working vacuum in the shortest possible time. Probably it would be advantageous if thermal contact between specimen and sink could be decreased or broken during the etching procedure to reduce undesirable heating of the cold sink.

[1] Olsen, Smith, and Crittenden, J. App. Phys. 16, 425 (1945).
[2] L. B. Loeb, *Kinetic Theory of Gases* (McGraw-Hill Book Company, Inc., New York, 1937), p. 96.
[3] *Handbook of Chemistry and Physics* (Chemical Rubber Publishing Company, Cleveland, 1945), 29th edition.
[4] J. C. Slater, *Introduction to Chemical Physics* (McGraw-Hill Book Company, Inc., New York, 1939), p. 176.

Plasma Oscillator

GOTTFRIED WEHNER,
Wright Field, Dayton, Ohio
August 15, 1949

EXPERIMENTS with a stabilized low pressure arc discharge[1] resulted in the discover of a new kind of r-f oscillator. This oscillator is similar to a klystron, whose resonance circuits are formed inside the plasma by thin oscillation layers.[2]

Cathode is either a mercury arc cathode with auxiliary anode or an oxide cathode. The grid completely separates the anode space from the cathode space and has a mesh width (depending on the gas pressure), which makes it possible to stabilize the discharge between fired and unfired operation.[3] The increased voltage drop of such a stabilized discharge is concentrated near the grid.[4] Electrons reaching this region from the cathode side are accelerated into the anode space and form a beam of fast electrons, provided the gas pressure is low (<5 micron) and the mean free path is large compared to the electrode distances.

When another electrode (a plate facing the grid and called repeller) is inserted in the beam and the anode arranged outside the beam, oscillations arise with a well-defined frequency, which are tunable between 1000 and 4000 mc by changing anode voltage and anode current. The grid and repeller (both near cathode potential) are covered with dark sheaths between which the bright anode plasma at high potential is located.

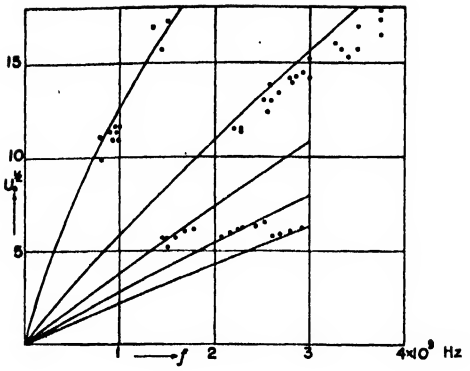


FIG. 1. Velocity *vs.* frequency diagram. Pressure 6 micron.

The investigation of the beam velocity distribution by means of a movable probe resulted in the discovery of a thin r-f velocity modulating layer inside the plasma near the grid ion sheath.[6] Often a second layer is found near the repeller ion sheath. In a suitable arrangement these oscillation layers can be seen as slightly darker zones inside the bright plasma.

Probe measurements showed that the plasma density N is not uniform along the anode space but decreases from the center in direction of the grid as well as of the repeller. Just near the oscillation layer N has a value which is connected with the observed frequency f by Langmuir's equation $f=9000N^{\frac{1}{2}}$. This density distribution provides an infinite number of resonance circuits in line with increasing resonance frequencies to the center. Hence only in one or two layers does the plasma density fit to one frequency. The repeller provides either the second oscillation layer (catcher) or the reversion of the beam when the tube oscillates as a reflex klystron.

Figure 1 shows the beam velocity *vs.* frequency diagram of a sealed-off all metal tube with an electrode arrangement according to Fig. 2. Without changing or matching any resonance circuit,

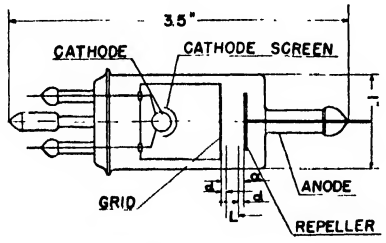


FIG. 2. Sealed-off plasma oscillator tube.

merely by changing anode voltage and anode current the frequency can be changed between 800 to 4000 mc. Oscillations occur along different modes. In case of higher pressures when reflex excitation due to the limited free path of the beam electrons does not exist, these modes are easy to calculate by transit time relations. Identifying the plasma edges near grid and repeller with the oscillation layers and assuming that the two oscillation layers oscillate in opposite phase, excitation should be possible if the transit time t_0 of electrons with the velocity v_0 through the drift space L equals:

$$t_0=\tfrac{1}{4}f\times(4n+1) \quad n=0,1,2,\cdots.$$

Setting

$$f=9000N^{\frac{1}{2}}$$
$$L=a-2d \text{ (Fig. 2)}$$
$$d=550U_0^{\frac{3}{4}}/N^{\frac{1}{2}}\,^{6,7}$$
$$v_0=6\ 10^7U_0^{\frac{1}{2}}$$
$$L=v_0t_0$$

the following relation may be derived:

$$(n+\tfrac{1}{4})\ 6\ 10^7U_0^{\frac{1}{2}}+10^7U_0^{\frac{3}{4}}=af.$$

This is shown with $a=0.6$ cm (Fig. 2) in the curves of Fig. 1. Beside these modes at lower pressures several other modes are detectable. They belong, as can be shown, to an excitation which is similar to the reflex klystron.

[1] W. O. Schumann, Archiv. f. Elektrotechnie 36, 362 (1942).
[2] H. J. Merril and H. W. Webb, Phys. Rev. 55, 1191 (1939).
[3] H. Fetz, Ann. d. Physik 37, 1 (1940).
[4] H. Fetz, Archiv. f. Elektrotechnie 36, 378 (1942).
[5] G. Wehner, Ann. d. Physik 40, 501 (1941).
[6] H. Fetz, Ann. d. Physik 40, 579 (1941).
[7] G. Wehner, Jahrb. d. deutsch. Luftfahrtforschg. III 24.

Wave-Lengths Suitable for X-Ray Microscopy of Biological Specimens

J. L. FARRANT
Division of Industrial Chemistry, Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia
September 6, 1949

AN x-ray microscope has recently been developed by Kirkpatrick and Baez[1] based on Compton's observation,[2] in 1922, that x-rays incident upon surfaces at glancing angles less than certain small critical angles undergo total specular reflection in accordance with the theory of Drude and Lorentz. The x-ray beam is focused by successive reflections at small glancing angles of incidence from two concave mirrors, having spherical, elliptical or cylindrical cross sections, and set up so that the normals to the two surfaces are mutually perpendicular. In this way one mirror exercises a focusing function for rays in a particular plane, while the other focuses the rays in a plane normal to the first. Kirkpatrick has suggested additional mirrors to reduce the spherical aberration and anamorphotism of the image inherent in the two-mirror system and calculates by means of the theory of Drude and Lorentz that diffraction limits the resolution to about 70A, this limit being independent of the wave-length of the radiation used.[1] In exploratory work with wave-lengths of the order of 1.5A at glancing angles of incidence of less than 30 minutes upon spherical mirrors coated with platinum and gold, Kirkpatrick and his colleagues have already obtained a resolving power of 2 microns.

Radiation of wave-length 1.5A has the advantage of being penetrating enough to allow the specimen to be imaged while in air at normal pressure, provided the absorption of radiation along the path between the mirrors and the image is diminished by the employment of a hydrogen or helium atmosphere.[1] On the other hand only the coarser details of comparatively thick specimens will absorb sufficiently to yield adequate image contrasts, since the "Half-Value Layer (H.V.L.) of organic substances (e.g., cellulose of density 1.3 g/cc) for a wave-length of 1.5A is approximately 0.7 mm, and even for an element of as high an atomic number as copper the H.V.L. is 15 microns. An additional limitation is that such wave-lengths restrict the angle of glancing incidence to a maximum of about 30 minutes, so that images are produced under conditions of extreme obliquity resulting in large aberrations.

In view of these limitations it is suggested by the writer that the realization of a microscope of this type suitable for the examination of bacteria and other organic specimens will entail the use of longer, more absorbable wave-lengths. Besides giving more adequate contrast the use of longer wave-lengths offers the additional advantages that (a) the degree of polish and accuracy of figure of the mirrors would be less critical, (b) the increased critical glancing angle of reflection would allow a greater numerical aperture with consequent increased speed of the optical system and (c) the larger aperture would allow more accurate focusing of the specimen with visible light before making the exposure with x-rays. It is also possible that conditions of less

TABLE I. Limits of resolution set by diffraction.

Mirror material	Wave-length (angstrom units)	$i_c \times 10^3$ (radians)	$d = \lambda/(2i_c)$ (angstrom units)	Observer
Platinum	2.0	14.1	70	E. Dershem[a]
Platinum	7.0	35.5	100	E. Dershem[a]
Gold	8.32	39.0	110	C. B. O. Mohr[b]
Glass	8.32	23.0	180	C. B. O. Mohr[b]
Glass	13.3	37.8	180	C. B. O. Mohr[b]
Glass	44.5	80.0	280	E. Dershem[c]

[a] E. Dershem, Phys. Rev. 33, 659 (1929).
[b] See reference 4.
[c] E. Dershem, Phys. Rev. 34, 1015 (1929).

oblique reflection would reduce aberrations, but more work on oblique optics is needed to decide this point.

According to the Drude and Lorentz theory the resolving power of the instrument is independent of the wave-length of the x-rays used.[1] However, the high absorption of long wave-length x-rays results in a less sharply defined critical angle and some of the radiation is specularly reflected at large glancing angles. Hence the Drude and Lorentz theory is inadequate here and even the application of the Fresnel expressions is not completely satisfactory. A closer agreement with experiment is given by the wave-mechanical treatment developed by Hönl,[3] but probably the best estimate of the resolution possible with long wave-lengths is obtained from actual experimental data, even though little information for monochromatic radiation has been published. In Table I is cited the limit d set to the resolution by diffraction, calculated by Kirkpatrick's formula $\lambda/2i_c$ from experimental values of the glancing angle of incidence i_c at which the reflected intensity I is half the intensity I_0 reflected at grazing incidence. The deterioration of resolution at long wave-lengths is not great.

Several investigators[4, 5] have found that mirrors composed of light elements are more efficient than heavy metals as reflectors of x-rays of wave-lengths greater than about 8A, and claims have been made for specular reflection of some of the radiation incident at surprisingly large glancing angles. The first observations of this nature were made by Holweck[6] in 1923, but perhaps the most interesting investigations are those of Laby, Bingham and Shearer,[7, 8] who examined the reflection of x-rays of wave-length 44.5A ($CK\alpha$) from several substances. At a glancing angle of 6 degrees soda-glass reflected 46 percent of the incident radiation and even at 20 degrees the reflected intensity was 18 percent. The radiation was focused with a concave soda-glass mirror at a glancing angle of incidence of 45 degrees to obtain a diffuse image of the focal spot. The observations made must be accepted with the reservation that part of the effect was due to general long wave-length radiation from the carbon anode employed, but a spectral analysis of the radiation revealed that it consisted largely of the carbon $K\alpha$-line. Although they realized their work established that concave gratings of considerable aperture could be used in this region, these workers unfortunately did not further explore the field of x-ray optics. These and other observations set out in Table II suggest that possibly mirrors of some materials would reflect sufficient of the very long wave-length radiations at close enough to normal incidence to allow the reflecting microscope of Burch[1] with its relatively large aperture to be used in that region.

TABLE II. Observations on the specular reflection of very long wave-length x-rays.

Anode material	Radiation X-ray tube potential (volts)	Mirror material	Glancing angle of incidence	I/I_0 percent	Observer
Molybdenum	150	Bronze	11°7′	35	F. Holweck[a]
Molybdenum	200	Bronze	16°2′	10	F. Holweck[a]
Nickel	194	Glass	20°	26	J. E. Henderson and E. R. Laird[b]
Nickel	576	Glass	20°	22	J. E. Henderson and E. R. Laird[b]
Molybdenum	130	Aluminum	15°	22	S. D. Gehman[c]
Molybdenum	150	Aluminum	12°	35	S. D. Gehman[c]
Molybdenum	300	Graphite	14°	22	S. D. Gehman[c]

[a] See reference 6.
[b] J. E. Henderson and E. R. Laird, Proc. Nat. Acad. Sci. 14, 773 (1928).
[c] S. D. Gehman, Phys. Rev. 33, 141 (1929).

Work on the specular reflection of x-rays seems to have practically ceased about 1932, but the attractions of x-ray optics are sufficient incentive to reopen investigation of this field. The modern techniques of vacuum evaporation and electron microscopy should greatly facilitate the preparation of mirrors and the examination of the smoothness of their surfaces.

An appreciation of the wave-lengths required to yield adequate image contrast for bacterial and organic specimens is given by the fact that the H.V.L.'s in organic substances for the $K\alpha$-radiations of aluminum (8.3A) and carbon (44.5A) are 5 and 1.4 microns respectively. The corresponding H.V.L.'s for air at N.T.P. are 4.5 and 1.0 mm, so probably the specimen could be imaged while placed in a thin cell filled with air or a lighter gas and sealed off from the x-ray tube and the optical system by thin, small diameter, vacuum-tight windows of collodion or beryllium. Radiation damage to specimens may however restrict the use of the longer wave-lengths.

The low intensity difficulty can be avoided to some extent by using visible light to search and roughly focus the field, while the photography of the image at low magnification with fine-grained plates should reduce exposure time somewhat. Application of the recently-developed x-ray image-intensity multiplier[9] may resolve this difficulty. The obvious advantages of the x-ray microscope are as follows: (i) Thicker specimens may be examined without the use of extremely high potentials. (ii) It may not be necessary to desiccate the specimen by placing it in vacuum. (iii) By choice of appropriate wave-lengths it should be possible to locate specific elements in the specimen. (iv) The x-ray microscope should be capable of use as an x-ray diffraction camera yielding focused diffraction patterns which would be of considerable value, particularly in the investigation of long spacings.

[1] P. Kirkpatrick and A. V. Baez, J. Opt. Soc. Am. 38, 766 (1948).
[2] A. H. Compton, Phys. Rev. 20, 84 (1922).
[3] H. Hönl, Ann. d. Physik 18, 42 and 625 (1933).
[4] C. B. O. Mohr, Proc. Roy. Soc. A133, 292 (1931).
[5] T. H. Osgood, Rev. Mod. Phys. 1, 228 (1929).
[6] F. Holweck, Comptes Rendus 176, 570 (1923).
[7] Laby, Bingham, and Shearer, Nature 122, 96 (1928).
[8] T. H. Laby and R. T. W. Bingham, Proc. Roy. Soc. A133, 274 (1931).
[9] Westinghouse X-Ray Division, Rev. Sci. Inst. 19, 477 (1948).

Thin Films Peeled from Teflon

C. K. JONES
Thomson Laboratory, General Electric Company, West Lynn, Massachusetts

July 28, 1949

THERE are many occasions when one has need of thin films of specific materials. These are cometimes difficult to prepare. The method described below was used to prepare films for x-ray absorption measurements.

A thin film of a barium compound was required with a negligible amount of foreign matter of high atomic weight. The specific compounds used were $BaCO_3$, BaF_2 and $BaSO_4$ but the method is equally applicable for a wide variety of such compounds.

The compound in question was mixed with General Electric Sleeving Lacquer #1548 in the ratio of one part compound to two parts lacquer by weight in a ball machine for two hours. A strip of Teflon, a vinyl polymer of the formula

$$\left[\begin{array}{cc} F & F \\ | & | \\ -C- & -C- \\ | & | \\ F & F \end{array}\right]_n,$$

which is attached only by liquid sodium, two inches wide, six inches long, and forty mils thick was mounted on a brass frame. The Teflon was then dipped into the mixture and withdrawn at a slow rate, the optimum being such that the rate of run back is equal to the rate of withdrawal. A Fisher-Payne dip coater is very useful for this purpose. The films were allowed to dry and

then peeled off. In removing the film, lifting at one end was found to be preferable to pulling the film back over itself.

Films prepared in this manner contained sixty percent compound, were as thin as ½-mil thick, and were sufficiently strong to be handled with ease. The thickness of the film could be varied by the dilution of the lacquer. The films were found to vary over a film and between films of roughly one percent.

High Strength Ceramics for High Temperature Use

RESULTS of recent tests at the National Bureau of Standards have shown that several ceramic bodies previously developed by the Bureau have marked superiority, in both strength and creep characteristics at 1800°F and above, over the best available high temperature metal alloys. Laboratory data indicate that these ceramic bodies, especially designed for use in jet engines and gas turbines, possess special properties required at the elevated operating temperatures of these power plants.[1]

The revolutionary developments during the past 30 years, particularly the last 10 years, in both design and efficiency of power plants have emphasized the need for materials of great strength and durability at very high temperatures. This is a natural consequence of the fact that, in any device for the conversion of heat energy into work, the efficiency of that device may be increased by increasing the temperature differential between the beginning and end of the conversion.

Because design engineers projected their plans into regions of temperature and stress far beyond the potentialities of known metallic alloys, a survey of non-metallic compounds, especially the oxides, silicates, carbides, and related combinations peculiar to ceramics became necessary. Little data existed in the literature upon which to base specific designs for the application of ceramics in power plants such as the gas turbine. Some work has been done on feldspathic bodies, which contain a bond of glass and, consequently, have limited use at elevated temperatures. Several reports were also available in the German literature on small specimens of simple oxide bodies with extraordinary resistance to mechanical stress at temperatures above 1800°F. These bodies had been matured, experimentally, at such high temperatures (about 3450° to 3550°F) that their successful production commercially in the United States seemed to be only a remote possibility.

Therefore, when the National Bureau of Standards was requested, by the National Advisory Committee for Aeronautics in 1944, to obtain engineering data on ceramic bodies in tension at elevated temperatures, it was thought advisable to begin the work on bodies which could be produced with available industrial facilities. Leading ceramic concerns were requested to submit specimens for preliminary tests of bending at elevated temperatures and of resistance to thermal shock. Also, six promising bodies were selected from a number developed at the National Bureau of Standards during a previous investigation. These tests, involving bending strength at 1800°F and relative resistance to thermal shock, showed the oxide bodies developed at the Bureau to have superior quality over those then available from commercial concerns. A comprehensive investigation was begun on the high temperature strength and creep in pure tension of six NBS oxide bodies.

For test purposes specimens round in cross section and approximately 0.3 in. in diameter throughout the gauge length were prepared. The adapters for applying the stress were 8½ in. long, 1 5/16 in. in diameter, and were made of one of the bodies under test. Ceramic adapters were used because calculations based on the data available indicated that high temperature alloys would not have sufficient strength at the maximum temperatures and stresses likely to be obtained in some of the tests. Both adapters and test specimens were fabricated under contract by a co-operating industrial concern, using hydrostatic pressure applied by means of rubber molds. The strain gauges, similar in design to those used at the National Bureau of Standards for measuring creep of metals, were fabricated from 90 percent platinum, 10 percent rhodium tubing, and wire.

TABLE I. *Stress-temperature data on four high strength ceramic bodies.*

Temperature (°F)	Maximum stress lb./in.² NBS ceramic bodies No. 358	No. 353	No. 151	No. 4811C
1500	13,000	13,000	12,000	14,000
1700	14,000	13,000	13,000	14,000
1800	17,000	18,000	18,000	18,000
1900	8000	4000	15,000	16,000
2100	...	...	...	6000

Four thermocouples mounted in the vicinity of the specimen provided temperature measurements at the ends and at the center of the gauge length and actuated the temperature controlling and recording instruments. The specimens were heated in conventional wire-wound furnaces equipped with viewing windows to permit the periodic measurement of strain using a Gaertner extensometer-viewing microscope. Loads were applied to the specimen through a lever system.

Four general methods of test were followed, the first two of which may be described as "step testing."

Method 1.—The specimen was heated gradually to the test temperature and a stress considered to be well below the tensile strength of the specimen was applied. After observing length changes for some predetermined time, the stress was increased. Length changes were observed again, whereupon another increment of stress was applied. This was repeated until rupture occurred. It was customary to increase the stress in increments of 1000 lb./in.² at about one-week intervals.

Method 2.—The original temperature and load conditions were essentially as described for Method 1. In this method, however, the load remained constant and the temperature was raised, usually in 100°F steps at about one-week intervals. In some cases, the specimen had not ruptured after prolonged holding at 1800°F. When this occurred in a furnace in which the temperature could not be raised safely above 1800°F, the test was continued by increasing the load in steps while maintaining the 1800°F temperature.

Method 3.—After holding the specimen at 1800°F under a stress of 4000 lb./in.² for 48 hr., stress was increased at the rate of about 1200 lb./in.²/min. to failure. The load was increased by flowing shot into a bucket.

Method 4.—The specimen was held at constant stress and constant temperature to failure, the purpose being to obtain curves of stress against rupture time which could be compared with similar curves for metal alloys.

In general, each combination of stress and temperature was maintained for about 160 hr., or one week. Conditions were then changed by an increment of either stress or temperature, until failure, for a total of 109 tests. Strengths in tension up to 18,000 lb./in.² at 1800°F, and 15,000 lb./in.² at 1900°F were observed. Above 1900°F, however, the strengths drop off rapidly to an average of about 5000 lb./in.² at 2000° to 2200°F.

Resistance to creep also decreases rapidly, for the particular porcelains tested, when temperatures are raised above the range 1800° to 1900°F. Results show that even the comparatively low stress of 6000 lb./in.² has caused more than a tenfold increase in the rate of creep at 2050°F compared to the rate at 1800°F,—and the porcelains on which these tests were made are the most resistant to creep of any investigated. Maximum observed creep rates, for all of the bodies tested, may be summarized as ranging from about 0.0001 to 0.0002 percent per hour at 1700°F for the range of stresses used; from 0.0002 to 0.0008 percent per hour at 1800°F and a stress of 16,000 lb./in.²; and from 0.0030 to 0.0040 percent per hour at 1900°F and 10,000 lb./in.² stress. The maximum stress at rupture for the four NBS ceramic bodies at various temperatures is given in Table I.

Data obtained from the Bureau's tests have revealed that some of the NBS oxide bodies are capable of effectively withstanding high temperature conditions far better than any other material now in use. In fact, an experimental gas turbine, using blades fabricated from the oxide body of the most promising strength and creep characteristics, has been operated successfully in the Cleveland Laboratory of the National Advisory Committee for Aeronautics.

[1] For complete technical details see Burdick, Moreland, and Geller, "Strength and creep characteristics of ceramic bodies at elevated temperatures," NACA Tech. Note No. 1561, available from the National Advisory Committee for Aeronautics, Washington 25, D. C.

PROGRAM

Proceedings of the Electron Microscope Society of America

Chairman: L. MARTON

THE annual meeting of the Electron Microscope Society of America was held at the National Bureau of Standards in Washington, D. C., October 6–8, 1949. Titles and abstracts of the papers presented are given below.

1. The Temperature of Electron Microscope Specimens. G. D. SCOTT AND FRANCES DAWES, *Department of Physics, University of Toronto, Toronto, Ontario.*—It is well known that with the high current densities of the illuminating beam obtainable with a self-biased gun, electron microscope specimens often undergo marked changes. Frequently these changes can be attributed to elevated temperatures. This paper is a discussion of the equilibrium between the energy absorbed by a particle from the electron beam and that which is either radiated to the walls of the surrounding chamber or conducted to the supporting substrate. For the case of an opaque particle on a Formvar film it can be shown that under possible conditions of contact between particle and film, conduction becomes negligible and hence the equilibrium temperature as determined by radiation alone can be 1000°C or more. As there is no mechanism by which the film can absorb any large fraction of the energy either directly from the beam or from the radiation emitted by the adjacent particle, the film itself remains at nearly ambient temperature. The high temperature gradients at the points of contact may, of course, cause instability and ultimately rupture of the film. The temperature conditions of electron microscope specimens are thus seen to be in marked contrast to those which can exist in macroscopic experiments.

2. Vibration Studies: Related to Electron Microscopy. F. A. HAMM AND F. C. SNOWDEN, *General Aniline and Film Corporation, Easton, Pennsylvania.*—This investigation has provided information which is useful as a convenient guide for properly locating the instrument and for high resolution electron microscopy where vibration is a problem. The scheme involves the measurement of an oscillographic trace whose amplitude is proportional to the voltage generated by a piezoelectric crystal pick-up. The essential steps in the calibration and a brief description of the circuit characteristics are given. Electron micrographs relating image shifts with mechanical displacements of the instrument are illustrated. At present these results are only semiquantitative.

3. Performance Tests on the RCA Model EMC Electron Microscope. S. G. ELLIS, *RCA Laboratories, Princeton, New Jersey.*—A standard EMC electron microscope was tested to find the factors which limited its performance. These factors were drifting of the specimen due to thermal motions of the specimen stage and objective lens asymmetry. When these had been reduced, the instrument demonstrated a resolution of approximately 50A. The resolution was now limited by the angular aperture of the illumination. Some of the techniques employed will be described.

4. On the Testing of Electron Microscope Objectives. JAMES HILLIER AND HENRY FROULA,* *RCA Laboratories, Princeton, New Jersey.*—A method of obtaining a quantitative evaluation of electron microscope objectives is described. The axial distance between the line foci produced by an asymmetrical objective is used as a measure of its asymmetries. Using a carefully compensated objective and a carbon black specimen (filmless mount), a photographic chart has been assembled showing the Fresnel fringes in images defocused by known amounts in the range $+100\mu$ to -100μ. These were obtained under standardized conditions of operating of the microscope. To test a lens a single micrograph of a similar specimen is taken near focus and under approximately the same conditions and an enlargement made at the same magnification as the chart. The fringes in the directions of maximum and minimum defocusing are then compared with those of the chart. In this way the amount of defocusing in the two directions can be accurately estimated. The difference between these values is the axial separation of the line foci for the lens tested. It is shown that a separation of $\frac{1}{2}\mu$ can be detected. Preliminary estimates indicate that separations of less than 1μ are necessary for the highest resolution. Lenses with separations greater than $10–15\mu$ are what would be generally considered as unsatisfactory.

* Permanent address: Department of Engineering, University of California, Los Angeles, California.

5. A Technique for the Comparison of Identical Surface Areas in the Light and Electron Microscopes. ERNEST F. FULLAM AND RUTH H. PETCHER, *Knolls Laboratory, General Electric Company, Schenectady, New York.*—To facilitate the interpretation of electron micrographs of surfaces, particularly metals, it is desirable to compare identical areas, at the same magnification, in the light and electron microscopes. The principle difficulty involved is in locating the identical area of the sample in each instrument. This has been done quite successfully by Bryner[1]; however, a simpler and more rapid technique would be more desirable. The proposed simplified technique consists of polishing and etching the metal sample in the usual way. Immediately following the etch an acute angle is scribed lightly on the surface, under the light microscope with a sharpened needle, enclosing or adjacent to the desired area. The needle is placed on the sample by an arm attached to the fine focusing mechanism from another microscope. This insures a very light contact with the metal to minimize the production of a cold-worked structure. The line width is three to ten microns. The angle is reproduced in a nitrocellulose replica cast on and stripped from the surface. This first replica is coated with a thin double film of beryllium and quartz. Quartz films alone do not have sufficient strength. A "Lektromesh" specimen screen with a four-mil hole drilled in its approximate center and polished is cemented to the replica with the apex of the angle visible in the center of the hole. The hole is necessary in order to locate quickly the angle in the electron microscope and to provide sufficient identifiable area adjacent to it for study without the interference of the screen wires. This placing operation is carried out under a dissecting microscope through which the angle in the replica is easily visible. After washing away the nitrocellulose replica by condensation of amyl acetate vapors on it, the remaining thin replica film over the drilled screen is ready for the microscope. Using the apex of the angle as a reference point, many areas adjacent to it may be easily recognized in both the light and electron microscopes even though the surface has no distinctive natural structure that may be readily identified.

[1] J. S. Bryner, "Correlation of optical and electron microscopy," Metals Tech. 15, No. 4 (June 1948).

6. A Centerable Condenser Aperture for the RCA Model "B" Electron Microscope. F. F. MOREHEAD, *American Viscose Corporation, Marcus Hook, Pennsylvania.*—A modification of the regular condenser aperture that may be centered while the instrument is in operation is described. (5 min.)

7. Permanent Magnet Lenses. JOHN H. REISNER, *RCA Victor Division, Radio Corporation of America, Camden, New Jersey.*—Permanent magnets have been successfully utilized to energize magnetic lens pole pieces for high resolution electron microscopy at direct magnifications of 7000 times with 30-kilovolt electrons, and 5000 times at 50 kilovolts. Stray magnetic fields which have previously limited high magnification and high resolution have been minimized by using two circuit gaps in parallel, and by surrounding the assembly with a magnetic shield connected to the external pole pieces of the two gaps. Magnetomotive forces of 1400 gilberts have been obtained. A single magnet or magnets in bucking arrangement, as well as configurations involving magnets disposed either radially or parallel to the optical axis of the lens, have been successfully utilized. The effects of various magnetic circuit parameters on the design and operation of the lens system have been studied.

8. Beam Deflection Focusing Device. HAROLD T. MERYMAN, *Naval Medical Research Institute, Bethesda, Maryland.*—A beam deflection device operating on an electromagnetic principle has been adapted to the RCA EMU electron microscope. Design and principle of the device will be discussed and its performance demonstrated with micrographs.

9. Motion Pictures of Electron Microscope Images. JOHN H. L. WATSON AND LUTHER E. PREUSS, *Edsel B. Ford Institute for Medical Research, Detroit 2, Michigan.*—The feasibility of taking motion pictures of electron microscope images has been considered on occasion, but the lack of informative, moving subjects has made the project impractical. Lately the technique has been developed as a tool for studying the effects of electron bombardment upon matter[1] and particularly upon tungsten oxide[2] and other tactoid structures. It may be used also to study reticulation of thin metal films under electron bombardment, contamination of microscope specimens due to the electron beam, and for observing the effects of bombardment upon biological and bacteriological specimens. The technique is described and several films are shown illustrating the results. It is found that intensity in the microscope is quite adequate and that contrast is good. Resolution is inferior to stills taken in the normal way, but images are interpreted from "moving micrographs," in conjunction with stills, somewhat more easily by virtue of (1) object movement which often gives a three-dimensional impression, (2) the control of the film which allows the reactions to be observed at a range of speeds and to be studied at close hand repeatedly, and (3) the increased number of single micrographs which are taken in a run and which, because of their number and the speed with which they are taken, are likely to contain more complete information than are a few isolated stills. (20 min.)

[1] John H. L. Watson, J. App. Phys. 19, 713 (1948).
[2] Heller, Wilfried, Wojtowicz, Wesley, and Watson, J. Chem. Phys. 16, 998 (1948).

Papers on Sectioning

Chairman: T. F. ANDERSON

10. Improved Sectioning Technique for the Electron Microscope. RICHARD F. BAKER AND DANIEL C. PEASE, *School of Medicine, The University of Southern California, Los Angeles, California.*—Some refinements in a previously described technique[1,2] are described. These include a reliable method of flattening and mounting sections, also a standardized method of knife sharpening. A comparison of fixing agents has been made, using liver and nervous tissue, which indicates the superiority of osmic acid for general use. Perfusion is compared with fluid and vapor immersion as a means of introducing the fixative to the tissue. A discussion of resolving power limitations in thin section microscopy is given. (20 min.)

[1] D. C. Pease and R. F. Baker, Proc. Soc. Exper. Biol. and Med. 67, 470 (1948).
[2] R. F. Baker and D. C. Pease, J. App. Phys. 20, 480 (1949).

11. Ultra-Microtomy by a New Method. SANFORD B. NEWMAN, EMIL BORYSKO, AND MAX SWERDLOW, *National Bureau of Standards, Washington, D. C.*—Polymerization of *n*-butyl methacrylate monomer is used as a rapid and simple means for embedding fixed biological material. The solid resin provides an optically clear matrix for cutting very thin sections, one at a time, with a slightly modified conventional rotary microtome. Advance of the embedded specimen toward the knife is obtained from the thermal expansion of a brass specimen holder. These ultra-thin sections have uniform thickness, large area, and integrity of tissue structure. They are suitable for producing transmission images at the higher magnifications of the conventional light, phase-contrast, and electron microscopes. Metallic shadowing of the sections provides greater contrast as well as a three-dimensional aspect to the structural details of the tissue. (20 min.)

12. Electron Microscopy of Renal Morphology. JOHN J. KELSCH AND JOAN BARDET, *Interchemical Corporation, New York, New York.*—The high speed microtome, developed at Interchemical Corporation Research Laboratories, and described by Fullam, Gessler, Grey, Kelsch, and Schuster has been redesigned. The knife is set in a rotatable holder to facilitate the adjustment of the critical clearance angle, which must vary for different materials. The wheel is heavier, and is balanced at three points. The knife edge is at a 45 degree angle to the radius of the wheel to give a slicing action in contrast to the chopping action of the previous model. Cutting speed has been reduced to *ca.* 10,000 r.p.m., and feed speed reduced to 0.5 mm/min. Utilizing the improved high speed microtome, a study of the renal morphology of the mouse was undertaken. The structure of the renal tubules as seen in 0.1μ-sections in the electron microscope is demonstrated, and a series of sections through the cortex and medulla is shown.

13. Some Refinements of the Rotary Microtome Modified for Ultra-Thin Sectioning.* JAMES HILLIER, *RCA Laboratories, Princeton, New Jersey* AND MARK GETTNER, *Sloan-Kettering Institute, New York, New York.*—Using the reservoir described in the accompanying paper accidental distortions of the section by the blade are eliminated with the exception of uniform compression. This appears to be a function of the sharpness of the blade and the thickness of the section cut. Thus with a given blade the compression of each section, which can be easily observed through a low power microscope, provides a reliable indication of the thickness. By this means it has been possible to identify many sources of variations in the thickness of the sections. In the Spencer rotary microtome static friction in the horizontal ways is largely eliminated by the introduction of a 300-micron horizontal excursion of the specimen block during each cutting cycle. This motion must be reproduced to within 0.01 micron if sections of uniform thickness in the 0.05- to 0.1-micron range are to be cut. Static friction in the horizontal ways which has an opportunity to occur during two extremely short intervals in each cutting cycle and is one of the major causes of thickness variation has been reduced by a spring which supports most of the weight of the horizontal member. Variations in the thickness of the

lubricating film in the rear vertical way and on the advance plate constitute a second important source of thickness variations. A definite improvement in performance has been obtained by keeping the thickness of this film to a minimum and by running the mechanism more or less continuously for one-half to one hour before cutting. The third source of variations that has been identified includes vibrations originating in the rachet and other parts of the advance mechanism, from irregularities of hand operation and from accidental table and room vibrations. Laying a finger lightly on any part of the microtome will cause as much as a 0.05-micron variation. The addition of a motor drive has so improved the operation that the variation in 0.02-micron sections cannot be detected visually. It is presumed they are below 40A.

* This work was sponsored in part by the U. S. Navy under Contract N6-ori-99 Task Order I and in part by RCA Laboratories.

14. Serial Sections for Electron Micrography.* MARK GETTNER, *Sloan-Kettering Institute, New York, New York,* AND JAMES HILLIER, *RCA Laboratories, Princeton, New Jersey.*—When a rotary microtome is modified[1] to cut sections of thickness less than 0.2 micron, serial sections cannot be cut in the usual fashion. The extremely low rigidity of such thin sections is not sufficient to overcome the friction between section and the facet of the knife resulting in a rolling and folding of the section. In the present work this has been overcome by the use of a reservoir attached to the blade and filled with a liquid. The height and surface tension of the liquid are adjusted until the meniscus reaches just to the cutting edge. The block and the clearance facet of the blade are kept dry. With this arrangement the sections float as they are being cut, the trailing edge of each section holding to the edge of the blade until the leading edge of the succeeding section attaches itself to it and pushes it on. In this way a chain of sections already flattened, spread and in sequence is obtained floating on the surface of the liquid from which it can be lifted and mounted by any of the usual techniques. The only distortion of the sections observed in this technique is a uniform compression which occurs when sections of less than 0.05 micron are cut or when the blade is dull. The use of different liquids and cutting speeds will be discussed.

* This work was sponsored in part by the U. S. Navy under Contract N6-ori-99, Task Order I and in part by RCA Laboratories.

[1] D. C. Pease and R. F. Baker, Proc. Soc. Exper. Biol. and Med. 67, 470 (1948).

Chairman: C. M. SCHWARTZ

15. A Technique for the Preparation of Grease or Solid Samples Dispersed in Grease-Like Media for Examination with the Electron Microscope. A. Y. MOTTLAU, *Standard Oil Development Company, Linden, New Jersey.*—The nature of grease as a dispersion of solid particles in oil is briefly discussed, and the value of the electron microscope in studying this solid phase is pointed out. The requirements of a satisfactory grease specimen mounting technique are outlined and a new technique fulfilling these requirements is described. Electron micrographs are presented which demonstrate the usefulness of the technique in disclosing pertinent information on grease structure. The factors involved in the preparation of powder specimens for electron micrography are discussed in relation to the adoption of a modification of the grease mounting technique for this purpose. The modified technique for powders is outlined and illustrated with micrographs.

16. Replication of Frozen Liquids by Vacuum Evaporation. HAROLD T. MERYMAN, *Naval Medical Research Institute, Bethesda, Maryland.*—Silicon monoxide replicas of frozen liquids have been made with particular attention to water. The size and shape of single crystals of ice frozen at various speeds can be demonstrated. The ices have been studied by replication of the surface immediately following freezing, following evaporation from the surface, and of the interior of a fractured crystal. The method has been applied to replication of wet biological specimens and its advantages and limitations will be discussed.

17. Measurement of Cellulose Particle Length by the Electron Microscope. F. F. MOREHEAD, *American Viscose Corporation, Marcus Hook, Pennsylvania.*—The technique for preparing cellulose films from viscose thin enough to examine directly in the electron microscope is described. A difference in particle size on the two sides of cellulose film cast as described on glass is reported and an explanation offered. The average particle length when corrected proportionally to the "skin" and "core" of two types of viscose rayon is compared with the lengths obtained from the measurement of "crystallites" observed after acid hydrolysis plus ultrasonic disintegration. Good agreement is found when the latter are corrected for recrystallization during hydrolysis. The length of the crystalline and amorphous areas of a number of cellulose fibers are calculated from the lengths of the hydrolysis crystallites as measured by the electron microscope. From the calculated results the ratio of the crystalline and amorphous lengths and the number of crystalline plus amorphous lengths per chain is obtained. (15 min.)

18. Reproducibly Representative Specimen Fields in Electron Microscopy. ROBERT C. BACKUS AND ROBLEY C. WILLIAMS, *University of Michigan, Ann Arbor, Michigan.*—Representative specimen fields are highly desirable in electron microscopy whenever one attempts to do precise qualitative or quantitative work. Developments leading toward such fields have been made in the past, but they have not been extensive, nor has there apparently been any adequate appreciation of the unique usefulness of representative fields. The method of producing representative fields by spraying has been extensively developed, and its advantages investigated. The qualitative advantages are the short drying time of spray droplets and the establishment of ideal conditions for both a precise qualitative assay of a single specimen and for a comparison of two or more specimens. The primary quantitative advantage is that counts can be made of the number of specimen particles per unit volume of suspension, by the use of reference particles. The suspending medium should be wholly volatile, in order to realize these advantages, and a volatile, electrolyte-containing buffer has been developed and tested.

19. Determination of the Weight of the Bushy Stunt Virus Particle by Electron Microscopy. ROBLEY C. WILLIAMS AND ROBERT C. BACKUS, *University of Michigan, Ann Arbor, Michigan.*—A method has been developed for determining the weight of macromolecules by counting them directly in electron micrographs. This paper describes the method and the results for the case of the bushy stunt virus particle. The number of virus particles per unit volume of a suspension is determined by counting the virus particles mixed with a known concentration of reference particles, as they appear in representative fields formed by the method described in the preceding abstract. The dry weight of the virus particle in a known volume of the suspension is measured, and hence the weight of a known number of particles is ascertained. Simple division yields the weight of a single particle. In the case of a highly purified suspension of the bushy stunt virus this method has yielded a particle weight of $9.4 \pm 0.7 \times 10^6$.

20. Electron Metallography of Cast Irons. CHARLES H. GEROULD, *The Dow Chemical Company, Midland, Michigan.*—A preliminary report is given on the metallography of cast

irons with emphasis on the formation and distribution of graphite. Electron micrographs are shown comparing typical normal cast iron with malleabilized, magnesium-treated, magnesium-overtreated, and cerium-treated cast irons showing graphite distribution and its relation to properties. Polystyrene-SiO replicas have been used in this work after the etched metal surface has been cleaned by stripping collodion from the surface.

21. **Quantitative Metallography with the Electron Microscope.** ALFRED L. ELLIS AND F. K. IVERSON, *International Harvester Company, Chicago, Illinois.*—An investigation of the carbide particle size and distribution in five heats of annealed SAE 52100 steel revealed that between 50 and 75 percent of the carbide particles are smaller than 0.5 micron. Since carbide particles smaller than 0.5 micron are difficult to measure at magnifications normally used for lineal analysis (1000×), it was considered desirable to develop an accurate quantitative method using the electron microscope. The accuracy of the replica method was first checked by examining the same carbide particles in the light and electron microscopes. A method was then developed for determining the percentage of carbide from electron micrographs. A statistical analysis of this method revealed that ten plates of five exposures each are required for accuracy within ±0.5 percent of the mean with 95 percent probability. However, only two plates are required for accuracy within approximately ±1 percent of the mean with 65 percent probability. The method of sampling and the statistical data will be presented. The method was applied to determination of the percentage of carbide in heat-treated steels.

22. **Electron Metallography of Cemented Carbides.** W. L. GRUBE, *General Motors Corporation, Detroit, Michigan.*— In the evaluation of quality and in the correlation of metallurgical properties with the performance of cemented carbide tools various metallographic examinations are utilized. Light-microscopical examinations of this type of material furnish information concerning inherent porosity, bonding, grain size, and grain size distribution. Electron-microscopical examination of cemented carbides has extended the study of the above-listed properties and has also made possible a study of microstructure heretofore unresolved. The metallurgical properties of cemented carbides which have been studied by electron microscopy are: (1) differential etching of binder metal, tungsten carbide, and alloy carbide phases, (2) grain size (below two microns), (3) binder metal distribution, (4) segregation, (5) boundary phenomena between carbide grains, and (6) submicroscopic voids. The electron microstructure of specific samples of commercial carbide tools and its relationship to tool performance is discussed. The replica techniques used in the course of the examinations are also presented.

23. **Electron Diffraction Studies of Thin Iron and Chromium Films.** LELAND L. ANTES, *The University of Texas, Austin, Texas.*—The formation of air compounds on vacuum-deposited iron and chromium films at room temperature is discussed. When the thickness of the iron films is close to 10A, the conversion of the iron is nearly complete in a few minutes. The reaction of chromium with air is quite similar. The electrical resistance change of thin chromium films is discussed. The talk will be illustrated by slides.

24. **Electron Microscopy of Enamel and Dentin.** DAVID B. SCOTT AND RALPH W. G. WYCKOFF, *National Institutes of Health, Bethesda, Maryland.*—The fine structure of enamel and dentin has been studied through use of shadowed collodion replicas of etched ground sections of teeth. Of special interest are the fibrous components of the dentin, which include the thin-walled dentinal fibers and their branches, fibers apparently within the dentinal fibers, collagenous fibrils, and ultra-fine fibers in the matrix.

25. **Studies of the Action of Sodium Fluoride on Human Enamel by Electron Microscopy and Electron Diffraction.** DAVID B. SCOTT AND RALPH W. G. WYCKOFF, *National Institutes of Health, Bethesda, Maryland,* AND ROBERT G. PICARD, *RCA Victor Division, Camden, New Jersey.*—Slabs of enamel, cut parallel to tooth surfaces, were immersed in 2–4 percent NaF solutions for periods ranging from 4 min. to 30 days. Under the electron microscope an inhomogeneous apparently crystalline deposit was observed on replicas taken after 15–30 days; traces were seen after 2–7 days; under 2 days results were negative. Electron diffraction indicated that the deposit is CaF_2; the normal apatite pattern was replaced by that of CaF_2 after treatment as short as 3 min. When treated slabs were washed for prolonged periods in water the deposit was not seen microscopically, and the diffraction pattern reverted to that of the normal apatite. (15 min.)

26. **The Molecular Configuration in Seed Globulin Crystals.** C. E. HALL, *Massachusetts Institute of Technology, Cambridge, Massachusetts.*—A method for examining the surface structure of small protein crystals will be described. Micrographs have been obtained showing the molecular configuration at the surface of crystalline edestin and certain other seed globulins. It is concluded from the observations that the structure is face-centered cubic with $a = 114$A in the dry state. The edestin molecule appears as an approximately spherical particle about 80A in diameter. Small-angle x-ray powder patterns made by Dr. R. S. Bear show spacings consistent with the conclusions drawn from the electron microscope analysis.

Invited Papers on Complementary Methods of Microscopy

Chairman: F. O. SCHMITT

27. **Recent Developments in Conventional Microscopic Equipment and Techniques.** A. H. BENNETT, *American Optical Company, Stamford, Connecticut* (40 min.)

28. **Recent Developments of the Optical Microscope.** DAVID S. GREY, *Polaroid Corporation, Cambridge, Massachusetts* (40 min.)

29. **An Approach to X-Ray Microscopy.** PAUL H. KIRKPATRICK, *Stanford University, Stanford, California* (40 min.)

30. **Some Comments on Electron Microscope Research in Europe.** E. G. RAMBERG, *RCA Laboratories, Princeton, New Jersey.* (30 min.)

Chairman: C. E. HALL

31. Electron Microscope Study of Isolated Nuclei of Liver Cells from Laboratory Animals. JUAN J. ANGULO, *University of Havana, School of Medicine, Havana, Cuba,* AND JOHN H. L. WATSON,* *Edsel B. Ford Institute for Medical Research, Detroit 2, Michigan.*—Nuclei of liver cells from the Rhesus monkey, guinea pig, rat, mouse and the 16-day chick embryo were isolated by fractional centrifugation and examined without chemical fixation in the electron microscope. Structural details are seen and the identity with nucleoli of certain dense zones in the images is discussed. A relationship is noted between the number of these zones and animal species. Some zones show images which suggest the existence of a nucleolus wall but no evidence is found for the existence of a nucleolus membrane. The absence of a chemical fixation does not distort the nucleus structure and electron microscopy of unfixed, isolated nuclei seems to be superior in some ways to other techniques for electron microscope study of the nucleus.

* To give the paper.

32. An Electron Microscope Study of an Undescribed Orchid Virus. NELSON NEWTON AND DAVID W. ROSBERG, *Battelle Memorial Institute, Columbus, Ohio.*—In connection with phytopathological research on a heretofore unreported virus which produces mottled streak symptoms in orchid plants, electron microscope studies of both diseased and healthy plants have been made. Host range studies have shown the orchid to be the only plant capable of being infected with the virus. Purified and concentrated extracts from raw plant juice were prepared by alternate low (6460 g) and high (19,942 g) speed centrifugation. The material was finally suspended in a 0.05 M phosphate buffer at *p*H 7. Electron microscope examinations of platinum-shadowed specimens prepared from diseased and apparently healthy plants showed both crooked and straight rod-shaped particles present in the diseased material as well as in some of the apparently healthy extracts. Extracts from other apparently healthy plants were free of these rod-shaped particles. The straight rods were approximately 300×24 mμ, while the crooked rods were 400×21 mμ. This suggests that either two viruses or two different phases of the same virus were present. Buffer suspensions of the rod-shaped particles are opalescent and exhibit pronounced Tyndall cone phenomena. When examined with the polarizing microscope, intense streaming birefringence has been observed, comparable to that exhibited by tobacco mosaic virus.

33. Destruction of Bacterial Viruses by Osmotic Shock.* THOMAS F. ANDERSON, *Johnson Foundation, University of Pennsylvania, Philadelphia 4, Pennsylvania.*—The similar viruses, T2, T4, and T6 which appear in the electron microscope to have membranes surrounding the internal structures of the heads can be disintegrated by what might be termed "osmotic shock." If these viruses are incubated with 4M NaCl and the suspension then rapidly diluted with water, the viruses are inactivated, while little inactivation occurs if the suspensions are gradually brought to lower solute concentrations. Electron micrographs of slowly diluted preparations reveal normal virus particles, while micrographs of the former, rapidly diluted preparations reveal virus "ghosts"—empty head membranes with tails attached. Similar destruction of these viruses follows rapid dilution from solutions of ethylene glycol, glycine, glycerol, glucose, or sucrose. These results are compatible with the view that the membranes of the even-numbered viruses, like those of cells, are somewhat permeable to these solutes, but much more permeable to water molecules. Presumably, the virus heads swell when the osmotic pressure is suddenly reduced, and actually burst if the reduction is sufficiently large and sudden. Osmotic shocks which inactivate the even-numbered viruses have no detectible effect on the small viruses T1, T3, and T7 which appear in the electron microscope not to have membranes.

* This work was supported by Contract N6-ori-168 T.O. II between the ONR and the Trustees of the University of Pennsylvania.

34. "Crystalline" Virus-Like Bodies from Human Skin Papillomas. JOSEPH L. MELNICK, MAURICE J. STRAUSS, HENRY BUNTING, AND ERNEST SHAW, *Yale University School of Medicine, New Haven, Connecticut.*—Certain skin papillomas have been shown to be characterized by the presence of intranuclear inclusion bodies. Immediately following removal from patients, such papillomas were fractionated by differential centrifugation and examined in the electron microscope. As control samples, common warts, molluscum contagiosum lesions, and normal skin were studied. From the intranuclear inclusion body papillomas, spherical bodies were readily obtained. These were characterized by their arrangement into crystalline-like clusters or layers in which the average particle diameter was 52 mμ. When these particles were not in crystalline array, they averaged 68 mμ in diameter. In the electron microscope the crystalline arrangement was not unlike that previously noted for certain of the plant viruses. Control preparations of molluscum contagiosum lesions revealed characteristic brick-shaped elementary bodies. The common wart and normal skin preparations revealed no uniform particles, but merely amorphous scattered clumps of matter, collagen fibers, and spherical particles of varying diameter.

35. Plant Virus Growth Curves as Determined by Use of the Electron Microscope. RUSSELL L. STEERE, *University of Michigan Ann Arbor, Michigan.*—Using uniform Dow latex polystyrene particles and applying the spray technique developed by Backus and Williams, we have been able to follow the increase in extractable virus content of plants following inoculation. The procedure involves clarification and dilution of the plant juices, addition of a constant amount of latex and thorough mixing, taking the electron micrographs, and obtaining the ratio of virus particles to latex particles in numerous drop patterns. This procedure is followed for each sample throughout the early development of the virus in the plant host. A comparison of the virus-latex ratio throughout the period results in what might be called a virus growth curve.

36. The Multiplication of Insect Viruses as Organisms. G. H. BERGOLD, *Laboratory of Insect Pathology, Sault Ste. Marie, Ontario.*—Electron micrographs of purified preparations of four different insect viruses indicate the presence of morphologically different forms, which are probably stages of multiplication. The virus first appears as a minute spherical body. This body increases in size and the virus appears as an elongated curved body surrounded by a membrane. Later the virus particle straightens out, ruptures the membrane and appears as a rod-shaped particle characteristic of insect viruses. One may assume that the rod-shaped virus particle contains several smaller sub-units each of which develops into a rod. The complicated nature of multiplication indicates that insect viruses are organisms with a relatively simple morphological structure of the mature rod.

37. The Electron Microscopy of Bacteriophage. RALPH W. G. WYCKOFF, *National Institutes of Health, Bethesda, Maryland.*—Several strains of the bacteriophages against colon bacilli have been under investigation to find out as much as possible about the mechanisms of their growth and multiplication. The electron microscopic results already obtained, to

be illustrated by a series of slides, show how these virus particles attack and destroy susceptible bacteria and make clear many steps in the process whereby they increase at the expense of these bacterial hosts. Reference is made to still incompletely understood aspects of bacteriophage proliferation, and to the bearing of these results on the more general problem of how viruses and other biological particles of macromolecular dimensions are produced in nature. (20 min.)

38. The Adaptation of Cytological Technique to Electron Microscopy. STUART MUDD AND ANDREW G. SMITH, *Medical School, University of Pennsylvania, Philadelphia, Pennsylvania.*—A limiting factor in the electron microscopic study of internal structures such as the nucleus in microbiological specimens has been inadequate contrast. Certain light cytological methods provide contrast but the light microscope has inadequate resolving power for dealing with the structure of the nucleus of bacteria. In the present study techniques of light cytology have been adapted to electron microscopy. The effect on contrast in the electron microscope of each step of the cytological processing has been determined. Technical details and the picture of the bacterial nucleus resulting will be presented.

39. The Effect of Certain Fixatives, Salts and Enzymes on Bacterial Cells with Respect to Chromatinic Structure and Contrast in Electron Microscopy. JON JONSEN AND ANDREW G. SMITH, *Medical School, University of Pennsylvania, Philadelphia, Pennsylvania.*—Bacterial cells which have been suitably fixed, treated with hydrochloric acid and stained show what has been referred to as chromatinic structure of a nuclear apparatus. Similar structures are produced after treatment with ribonuclease. In the present study the same structures were obtained with different salts on air-dried bacteria and bacteria fixed in Chabaud's fixative. Electron micrographs of similarly treated specimens will be shown and the results discussed together with a possible explanation of the salt effect. In addition, bacterial cells treated with osmium tetroxide vapor, Schaudinn's fixative, hydrochloric acid, desoxyribonuclease and trypsin will be presented.

40. A Fibrous Component of the Nerve Axon. FRANCIS O. SCHMITT, *Massachusetts Institute of Technology, Cambridge, Massachusetts.*—In a re-examination of the origin of the dense-edged fibrils (neurotubules) previously described by De-Robertis and Schmitt a study was made of giant fibers of the squid and the marine annelid, *Myxicola infundibulum.* Axons were isolated from the sheaths of formalin-fixed fibers and fragmented with a microblendor. Neurotubules were observed in the sheath of squid nerves from which the axoplasm had been removed before or after fixation; they were observed only occasionally in axon material (sheath contamination?). However, in some preparations no neurotubules could be found. More information is needed before the origin and composition of neurotubules can be stated conclusively. The most prominent feature of formalin-fixed axons is a fibrous component consisting of a reticulum of nodose fibrils 200–400A in width and of indefinite length. The nodes sometimes appear to be regularly spaced though a definitive periodicity has not been demonstrated. Squid axoplasm, extruded from freshly dissected fibers into water, disperses slowly, yielding fibrous structures ranging from relatively thick, rod-like bundles to very thin filaments. This fibrous protein, which may be concentrated by centrifuging at 10,000 g, is probably the chief fibrous component of the axon; its chemical properties are being studied. Even after brief blending, which fragments the fibrous bundles, diluted axoplasm yields nodose fibrils, after formalinization, which resemble closely those obtained from axons isolated from formalin-fixed fibers. The bearing of these results on the structure of the fresh axon will be discussed.

41. Morphologic Characteristics of the Oral Fusiform Bacilli as Revealed by the Electron Microscope. EDWARD G. HAMPP, DAVID B. SCOTT, AND RALPH W. G. WYCKOFF, *National Institutes of Health, Bethesda, Maryland.*—Pure cultures of seven strains of oral fusiform bacilli have been used. Shadowed preparations have been made from cultures of each strain at ages varying from 24 hours to 20 weeks. The morphologic characteristics of the organisms and the changes which occur with age have been studied.

Chairman: R. C. WILLIAMS

42. Use of Polystyrene Latex 580 G, Lot 3584 in Electron Microscopy. CHARLES H. GEROULD, *The Dow Chemical Company, Midland, Michigan.*—A review is presented of size determinations on this latex from 16 different laboratories, 12 of which show an average size of 2593A±40A. The four other determinations range upward to 3200A. Data are presented concerning latex dilution which indicates that particle diameter is not appreciably influenced by degree of dilution, size of drop added to specimen, drying speed, etc. Precautions in handling the latex are reviewed, including freezing, the danger of water evaporation from the undiluted latex, chemical and bacterial contamination, change of pH, over-heating in the electron beam, fusing due to insufficient dilution and contamination under electron bombardment. Electron micrographs are included which illustrate several of these dangers.

43. Particle Size and Density of Dow Latex 580 G—Measurements with the Ultracentrifuge. D. GORDON SHARP, *Duke University, Durham, North Carolina.*—Sedimentation velocity studies of the particles of Dow Latex 580 G suspended in physiological saline solution (0.9 percent NaCl) have been made in order that they may be used as a standard reference material in virus work. They sediment as a monodisperse system; less than 1 percent of the material differs ±2 percent from the mean sedimentation rate. The sedimentation constant is $S_{w20}^{\circ} = 1960 \times 10^{-13}$ sec.$^{-1}$ extrapolated to infinite dilution. Particle density was measured by observing sedimentation rate in media whose density was increased by addition of D_2O and of bovine serum albumin. Although some boundary instability was encountered near the point of zero sedimentation the particle density is estimated to be 1.054±.001, the same result being given by both albumin and D_2O indicating little or no hydration of the latex particles. This is in sharp contrast to the behavior of the heavily hydrated influenza virus, for example. From these data it is calculated that the particle size of Dow Latex 580 G is 252±2.5 mμ. It is therefore useful as a standard of size in electron microscopy and of sedimentation constant in the ultracentrifuge. It may also be used with the latter instrument to measure viscosity of samples too small to be measured otherwise.

44. The Apparent Size of Small Semi-Opaque Objects. RAYMOND A. KERN AND S. F. KERN, *Lilly Research Laboratories, Indianapolis, Indiana.*—It will be shown that the apparent size of small semi-opaque objects is a function of the conditions under which they are observed. Slides will be shown, demonstrating the apparent size differences of Dow's polystyrene balls 580 G, Lot 3584, as well as other small objects such as bacteriophage. A tentative explanation will be offered for these observed size variations.

45. A Removable Intermediate Lens for the RCA Model EMU. JAMES HILLIER, *RCA Laboratories, Princeton, New Jersey.*—An intermediate lens which can be introduced into the RCA Model EMU electron microscope without modifica-

tion of the column and with only a slight change in the electrical circuits is described. When used in conjunction with the highest power projection lens pole pieces, the new lens provides a full field at ten magnifications in the range 900 to 24,000. In addition, the new lens preserves the easy accessibility of the pole pieces for cleaning and adjustment. The distortion is less than with the single lens for all magnifications except the maximum and is very nearly zero for the lower magnifications. One value of the magnification is essentially independent of the accelerating potential and hence provides the possibility of achieving a permanent magnification calibration.

46. An Optical Method for the Thickness of Thin Films. G. D. SCOTT, *Department of Physics, University of Toronto, Toronto, Ontario.*—Tolansky* has discussed a method of measuring film thickness by multiple beam intereference using fringes of equal chromatic order in reflection. This method has proved very satisfactory for the accurate thickness determinations of films, evaporated metals and others, which are used in electron microscopy. The equipment consists essentially of a white light source and a spectroscope with calibrated wave-length scale. The preparation of the interference plates and the experimental arrangement will be discussed in detail. It has been found that ordinary microscope slides are superior to optically ground flats for the interference plates and their use makes it possible to reduce the fringe order to as low as *one* or *two*. The ease of observation and the accuracy of the measurements are thereby increased. The method can be applied to all types of stable films in thickness from 10A to several microns with areas of film less than one square mm. The measurements give the average metric thickness of the film and an accuracy of within ±5A is readily obtained.

* S. Tolansky, *Multiple Beam Interferometry* (Clarendon Press, New York, 1948).

47. A Technique for Studying the Electron Microstructure of Steel. D. M. TEAGUE AND L. H. ZIMMERMANN, *Chrysler Corporation, Detroit, Michigan.*—A brief history will be presented of the formation of the "Joint Committee on Electron Microstructure of Steel." The aims and scope of this group will be reviewed, together with sample and replica preparation methods. The second portion of the paper will be a discussion of a modified Formvar technique developed in our laboratories for studying the microstructure of steel. The salient features of the procedure are: (1) Deep etching of metal surface. This provides a strong Formvar film with good contrast. (2) Use of a soap monolayer on the metal surface to facilitate stripping. (3) Shadowing at a high angle—approximately 45°. This step improves detail, yet avoids certain difficulties of interpretation. The resultant electron micrographs correlate more readily with the experience of the metallurgist who will utilize the results. (15 min.)

48. Recent Advances in Electron-Microscopical Resinography. M. C. BOTTY, T. G. ROCHOW, AND F. G. ROWE, *Research Laboratories, American Cyanamid Company, Stamford, Connecticut.*—The general resinographic method of manifesting the electron-microscopical structure inside any solidified resin by replicating its fracture-surface was recently described.[1] The present paper will stress the description of the electronmicroscopical techniques and will exhibit micrographs of the *macromolecules* of single and composite resins as well as incorporations of pigments, fillers, and other modifiers.

[1] Anal. Chem. 21, 461-466 (1949).

49. Continuous Observations on the Formation of Metallic Films in the Electron Microscope. R. S. SENNETT AND T. A. MCLAUCHLAN, *University of Toronto, Toronto, Ontario.*—A number of different metals, including silver, gold, cadmium, and zinc, have been evaporated in a special specimen holder designed for the RCA type EMU instrument. Continuous observations have been made on the films during their formation, and micrographs of the same area have been obtained at a number of different stages during the evaporation. In all cases the first indication of a film was the appearance of a number of nuclei, and additional evaporation only increased the size of these aggregates without increasing their number. In zinc and cadmium films, the nuclei which first appeared had a regular crystalline form which was usually distorted during subsequent growth.

Journal
of
Applied Physics

Volume 21, Number 2 February, 1950

A Note on Collapsing Cylindrical Shells

THEODORE E. STERNE
Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland
(Received May 23, 1949)

Expressions are obtained for the instantaneous dependence of pressure upon radius within a collapsing cylindrical shell of incompressible liquid when subjected to (a) neither external nor internal pressure or to (b) an external pressure but no internal pressure. Dynamical considerations suggest that very high pressures and radial velocities should be reached.

I.

DYNAMICAL considerations suggest that if an undisturbed cylindrical shell of liquid contracts radially, the velocity of the outer surface should tend to diminish and of the inner surface to increase, and that very high velocities and pressures should be reached. The author thinks that the collapse of a cylindrical liquid shell may bear some relation to details of the behavior of cylindrical and possibly of conical metal shells, caused to collapse by the detonation of external charges of high explosive. The collapse of such shells has already been considered, hydrodynamically, by Birkhoff, MacDougall, Pugh, and Taylor[1] without regard to such details as the instantaneous dependence of pressure upon radial coordinate.

II.

It is felt that the following brief treatment, which has been discussed with Professor Garrett Birkhoff, of an ideal case in which the liquid is incompressible and the collapse radial, without axial motion, may be of sufficient interest to merit a somewhat wider distribution and that it may also perhaps be regarded as a first step toward a detailed theoretical study of the collapse of a metal shell surrounded by an explosive charge.

III.

Consider a long liquid cylindrical shell with external and internal radii r_2 and r_3, respectively, and mass m per unit length. The quantity $r_0^2=r_2^2-r_3^2$ will be conserved as the cylinder contracts if, as is supposed, the motion is radial and the liquid incompressible. Let r_2/r_0 be denoted by x. Then the kinetic energy per unit length is

$$T=(m/2)r_0^2x^2\dot{x}^2\log[x^2/(x^2-1)],$$

and this is conserved if the shell is set contracting radially and left undisturbed. Any element of matter in the shell can be labeled by a Lagrange hydrodynamical coordinate μ defined as the mass, per unit of length, contained between the outside of the shell and a coaxial cylinder through the element. The velocity of any element of matter characterized by a particular value of μ is

$$\frac{\partial}{\partial t}r(\mu,t)=-\left(\frac{2T}{m}\right)^{\frac{1}{2}}\frac{1}{[\log^{\frac{1}{2}}x^2/(x^2-1)][x^2-\mu/m]^{\frac{1}{2}}}.$$

It can be seen from the preceding equation that the velocity of the internal surface approaches infinity as x approaches unity, and that the velocity of every other element of matter in the shell approaches zero as x approaches unity. It follows that during the undisturbed radial contraction of a cylindrical shell, the kinetic energy per unit mass tends to be more and more strongly concentrated toward the center as the contraction proceeds.

IV.

It may readily be shown that the pressure, zero at the outer and inner surfaces, has a maximum at

$$\mu/m=x^2-x^2(x^2-1)\log x^2/(x^2-1).$$

[1] Birkhoff, MacDougall, Pugh, and Taylor, J. App. Phys. 19, 563 (1948).

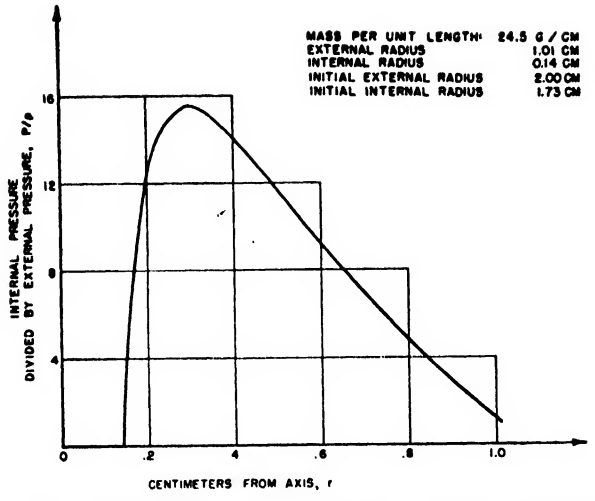


FIG. 1. The distribution of pressure in a radially contracting cylindrical shell of liquid, subjected to pressure over its outer surface.

The maximum pressure occurs at a place about midway between the surfaces when the shell is large, but moves toward the inner surface as the shell contracts. When the external and internal pressures are zero, the pressure is given generally throughout the cylindrical shell by the equation

$$P=\frac{1}{2\pi}\int_0^{\mu}\frac{1}{r}\frac{\partial^2 r}{\partial t^2}d\mu,$$

and has the value

$$P=\frac{T}{\pi r_0^2}\frac{1}{x^2(x^2-1)\log^2 x^2/(x^2-1)}\times\left[\log\frac{x^2}{x^2-\mu/m}-\frac{\mu}{m}\frac{x^2-1}{x^2-\mu/m}\log\frac{x^2}{x^2-1}\right].$$

Enormous pressures can be reached.

V.

If a constant pressure p is exerted on the external surface of the cylindrical shell, initially at rest, and zero pressure on the internal surface, then the resulting velocities and pressures can be calculated. The radial velocity of an element of matter in the shell is given by

$$\frac{\partial}{\partial t}r(\mu,t)=-\left(\frac{2\pi p}{m}\right)^{\frac{1}{2}}r_0\frac{(x_0^2-x^2)^{\frac{1}{2}}}{[\log x^2/(x^2-1)][x^2-\mu/m]^{\frac{1}{2}}},$$

where x_0 is the initial value of x, and the pressure in the shell is given by

$$P=p\left[1-\frac{\log x^2/(x^2-\mu/m)}{\log x^2/(x^2-1)}+\frac{x_0^2-x^2}{x^2(x^2-1)}\frac{\log x^2/(x^2-\mu/m)}{\log^2 x^2/(x^2-1)}-\frac{\mu}{m}\frac{x_0^2-x^2}{x^2(x^2-\mu/m)\log x^2/(x^2-1)}\right]$$

which has a maximum at

$$\frac{\mu}{m}=x^2-\frac{x^2(x^2-1)(x_0^2-x^2)\log x^2/(x^2-1)}{x_0^2-x^2-x^2(x^2-1)\log x^2/(x^2-1)}.$$

VI.

As an example of the application of the preceding general results, the following calculation may be of interest. If $p=10^{11}$ dynes/cm^2, if the initial value of the external radius r_2 is 2 cm, if r_0 is 1 cm, and if the liquid shell has the density of steel so that m is 24.5 g, then when $x=1.01$ the maximum pressure is at a value μ/m of 0.937 and is 15.6×10^{11} dynes/cm^2, which is sixteen times greater than the external pressure. The maximum velocity is then approximately 9.84×10^5 cm per sec., and is of course at the inner surface. In the preceding example the initial value of the internal radius is 1.73 cm and at the stage of contraction considered the external and internal radii are respectively 1.01 cm and 0.14 cm. The pressure is a maximum at 0.29 cm from the axis.

VII.

It seems intuitively plausible that the compressibility of any real liquid would probably cause the velocities and pressures to be less than they would be in an incompressible liquid. The elastic and plastic properties of a metallic shell would involve further departures from the ideal case of an incompressible liquid. The author is now unable to predict the magnitude of such departures. The magnitude may be limited, however, in cases of axially progressive collapse of the type considered by Birkhoff, MacDougall, Pugh, and Taylor.[1]

Steady-State Internal Temperature Rise in Magnet Coil Windings

CLAUDE L. EMMERICH*
Applied Science Research Laboratory, University of Cincinnati, Cincinnati, Ohio
(Received May 16, 1949)

The temperature distribution in an electric coil winding is determined by solving the differential equations which apply to the system approximately. The maximum and average temperature rises are given for several types of windings. A distinction is made between small and large temperature rise; the treatment of the former neglects local changes in resistivity and thermal conductivity caused by the increase in temperature. A correction factor is computed to obtain the actual temperature rise from the much simpler calculations for the maximum temperature rise which would occur if the resistance were independent of temperature. It is found that non-uniform windings, wound of several different wire sizes to minimize the resistance, exhibit the same temperature rise characteristics as conventional uniform windings of zero curvature.

INTRODUCTION

THE differential equation which applies to the steady-state temperature distribution in a magnet coil winding is Poisson's equation:

$$\nabla^2 T = -Q/K, \tag{1}$$

where T is the temperature at any point in the winding, Q is the rate with which heat is generated per unit volume per unit time, and K is the heat conducted per unit area per unit time per unit temperature gradient. The symbol ∇^2 denotes the Laplacian differential operator. In Cartesian coordinates, Eq. (1) would read

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}=-\frac{Q}{K}.$$

A coil winding is a heterogeneous structure, including metallic wire, surrounded by an insulating coating, and separated by insulating spaces. The term "insulating" refers qualitatively to both thermal and electrical properties. Thus, Eq. (1) as applied to magnet coils must represent a very complex system. However, for practical applications, it is sufficiently accurate to assume that the winding is a homogeneous body, with definite values of Q and K, averaged over the local discontinuities.

For a winding wound of one wire size, the value of Q is determined from the equation

$$Q=\rho A^2/(fL^2h^2), \tag{2}$$

where ρ = resistivity of the metal in the wire, ohm-inch; A = number of ampere-turns; f = space factor (cross-section area of metal divided by area of the winding Lh); L = length of winding, inch; h = thickness of winding, inch, $=b-a$ (see Fig. 1).

The value of K is not easy to predict from the physical constants of the coil. Moore[1] has made a theoretical investigation related to the problem of determining K; however, it appears that the theory has not yet been developed to give values of K more easily and accurately than experimental methods. Since K is an average of the thermal conductivities of the materials in the coil, it will be constant or variable according to the thermal characteristics of the wire and of the insulating materials of which the coil is composed. If the internal temperature rise in the winding is small, K may be assumed constant; and Q will also be nearly constant because the resistivity of the wire is not changed very much by a small change in the temperature.

Thus the solution of Eq. (1) is simplified considerably in the range of small temperature rises.

SMALL TEMPERATURE RISE

In particular, the small temperature rise conditions will be applied to a winding in the shape of a hollow cylinder (Fig. 1). If the winding is thin in comparison with its inside and outside diameters and its length, the curvature of the turns may be neglected, and the flow of heat may be assumed in the radial direction only, satisfying approximately the ordinary differential equation

$$(d^2T')/(dx^2)=-Q/K, \tag{3}$$

where x is the radial distance from the center line of the winding to the point whose temperature is T'. This equation will apply also approximately to a winding on a rectangular core.

Two sets of boundary conditions will be adopted as representing actual operating conditions of long magnet coil windings: (a) Core thermally insulated, outside surface at a fixed temperature T_s and (b) both core surface and outside surface at the same fixed temperature T_s.

These boundary conditions have been chosen in the interest of simplicity. It is evident that an actual coil does not operate under such conditions; however, to obtain a first approximation, some ideal cases should be considered. The deviations for actual coils may be judged from work reported in the literature.[2-4]

* Present address: 270 Harvard Street, Cambridge, Massachusetts.

[1] A. D. Moore, *Fundamentals of Electrical Design* (McGraw-Hill Book Company, Inc., New York, 1927), p. 95.

[2] G. L. Moses, Prod. Eng. **10**, 317 (1939).

[3] H. C. Stewart and L. C. Whitman, Trans. A.I.E.E. **63**, 763 (1944).

[4] H. M. Beede and B. M. Cain, Elec. Eng. **66**, 87 (1947).

Using boundary conditions (a) in the solution of Eq. (3), the temperature distribution in the winding is found to be

$$\Theta'(x)=T'(x)-T_s=(Q/2K)(b^2+2ax-x^2-2ab), \quad (4)$$

where Θ' is defined as the temperature rise above the surface temperature T_s. The maximum temperature rise in the winding will occur at the inside surface ($x=a$), where

$$\Theta_m'=\rho A^2/(2fKL^2). \quad (5)$$

Equation (5) is plotted in terms of the dimensionless variable $2fKL^2\Theta'/\rho A^2=1$ in Fig. 2.

A useful criterion of internal temperature rise is the average temperature rise, obtained from a measurement of the increase in total resistance of the winding. The average temperature is defined as the temperature to which the entire coil would have to be raised in order

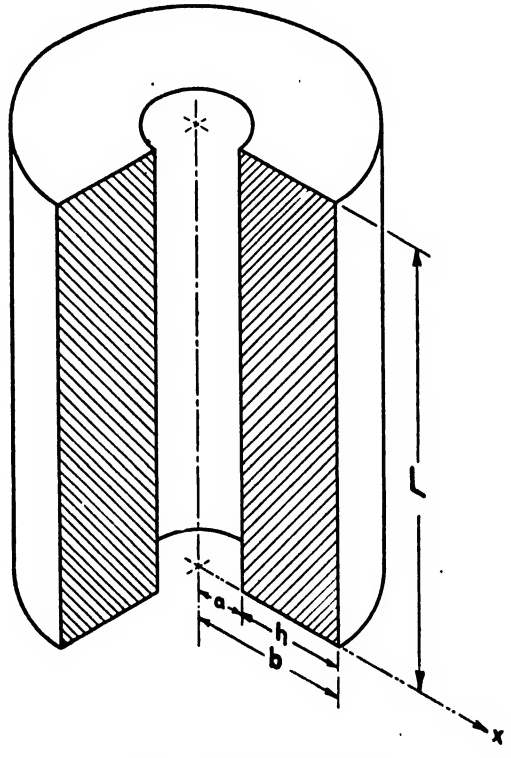


FIG. 1. Cylindrical winding.

to have the same increase in resistance as it actually has in the temperature field created by the heat generated in the winding. Expressed analytically, the average temperature rise is given by the equation

$$\Theta_{Av}=\left(\int\Theta dR\right)\Big/\left(\int dR\right), \quad (6)$$

where dR is the resistance of an elemental section of the winding whose temperature rise is Θ.

The temperature distribution as a function of the radius given by Eq. (4), together with the known distribution of resistance as a function of x, permits evaluation of the average temperature rise; which is found to be, for a winding with boundary conditions (a),

$$\Theta_{Av}'=\rho A^2/(3fKL^2). \quad (7)$$

This is plotted in Fig. 2 as $2fKL^2\Theta'/\rho A^2=\frac{2}{3}$.

For practical purposes it should be noted that the average and maximum temperature rises are in the ratio 2:3 for the thin cylindrical winding and small temperature rise.[5]

If boundary conditions (b) are applied to Eq. (3), the temperature distribution takes the form

$$\Theta'(x)=(Q/2K)(ax+bx-x^2-ab). \quad (8)$$

The maximum temperature rise will be the temperature rise at the center of the coil, where

$$\Theta_m'=\rho A^2/(8fKL^2). \quad (9)$$

In Fig. 2 this is plotted as $2fKL^2\Theta'/\rho A^2=\frac{1}{4}$.

The average temperature rise of the winding with boundary conditions (b) is

$$\Theta_{Av}'=\rho A^2/(12fKL^2) \quad (10)$$

plotted as $2fKL^2\Theta'/\rho A^2=\frac{1}{6}$ in Fig. 2.

For a thick, long, cylindrical winding, the assumption of negligible curvature, which led to the simple Eq. (3), cannot be applied. For this case, Poisson's Eq. (1) reduces to the ordinary differential equation

$$\frac{d^2T'}{dx^2}+\frac{1}{x}\frac{dT'}{dx}=-\frac{Q}{K}, \quad (11)$$

where x again denotes the radius of the point of temperature T'.

With boundary conditions (a), the temperature rise as a function of the radius is given by the expression

$$\Theta'(x)=(Q/4K)[b^2-x^2-2a^2\log_e(b/x)]. \quad (12)$$

The maximum temperature rise will occur at the inside of the winding, where

$$\Theta_m'=\frac{\rho A^2}{2fKL^2}\cdot\frac{r^2-1-2\log_e r}{2(r-1)^2}. \quad (13)$$

In this expression, $r=b/a$ is the thickness ratio of the winding, defined as the fraction outside radius divided by inside radius. Equation (13) will be found to represent actual conditions more accurately than the equation given for this case by Moore.†

Equation (13) may be represented by the following

[5] M. Jakob, Trans. A.S.M.E. 65, 593 (1943). Also M. Jakob and G. A. Hawkins, *Elements of Heat Transfer and Insulation* (John Wiley and Sons, Inc., New York, 1942), Chapter V.

† See reference 1, p. 98.

series expansion

$$\Theta_m' = \frac{\rho A^2}{2fKL^2}\left[\frac{(r+1)^2}{2(r^2+1)} - \frac{1}{3}(r-1)\left(\frac{r+1}{r^2+1}\right)^3 - \frac{1}{5}(r-1)^3\left(\frac{r+1}{r^2+1}\right)^5 - \cdots\right]. \quad (14)$$

The relation between $2fKL^2\Theta'/\rho A^2$ and r, as given by Eq. (13) is shown in Fig. 2 by the curve marked "Θ_m' uniform (a)." It is evident from this curve, as well as from the expansion (14), that for a thickness ratio of one, the special case of the thin winding is obtained. The other extreme, representing a winding in the shape of a solid cylinder, is seen to be approached as the thickness ratio becomes greater. For $r \to \infty$, Eqs. (13) and (14) reduce to the value

$$\Theta_m' = \tfrac{1}{2}\rho A^2/(2fKL^2). \quad (15)$$

which is just one-half of the expression for the maximum temperature rise of the thin winding [Eq. (5)]. It is indicated at the right edge of Fig. 2.

The average rise for the thick hollow cylinder with boundary conditions (a) can be determined from the temperature distribution $\Theta'(x)$ [Eq. (12)], and from the fact that the resistance of an elemental section of the hollow cylinder is proportional to the expression xdx. The average temperature rise is thus found to be

$$\Theta_{Av}' = \frac{\rho A^2}{2fKL^2}\cdot\frac{r^2-3+[4/(r^2-1)]\log_e r}{4(r-1)^2}. \quad (16)$$

Figure 2 shows this curve, marked "Θ_{Av}' uniform (a)."

The expansion of Eq. (16) is given by the following series

$$\Theta_{Av}' = \frac{\rho A^2}{2fKL^2}\cdot\frac{(r+1)^2}{(r^2+1)}\left[\frac{1}{4} + \frac{1}{3(r^2+1)^2} + \frac{(r^2-1)^2}{5(r^2+1)^4} + \cdots\right]. \quad (17)$$

If boundary conditions (b) are applied to the thick, long, hollow cylindrical winding [Eq. (11)], the temperature rise distribution will be found to be

$$\Theta'(x) = \frac{Q}{4K}\left(b^2 - x^2 - \frac{b^2-a^2}{\log_e(b/a)}\log_e\frac{b}{x}\right). \quad (18)$$

The radius x_m of the surface at which the maximum temperature rise occurs is found from the equation

$$x_m^2 = (b^2-a^2)/[2\log_e(b/a)].$$

The temperature rise at that surface will be

$$\Theta_m' = \frac{\rho A^2}{8fKL^2}\left\{\frac{2r^2}{(r-1)^2} - \frac{r+1}{(r-1)\log_e r}\left[1+\log_e\frac{2r^2\log_e r}{r^2-1}\right]\right\} \quad (19)$$

which is plotted in Fig. 2 as the curve marked "Θ_m' uniform (b)."

The average temperature rise with boundary conditions (b) is given by the equation

$$\Theta_{Av}' = \frac{\rho A^2}{2fKL^2}\cdot\frac{r^2+1-[(r^2-1)/\log_e r]}{4(r-1)^2} \quad (20)$$

shown in Fig. 2 by the curved marked "Θ_{Av}' uniform (b)."

In a non-uniform winding designed for minimum resistance,[6] the wire diameter varies as the square root

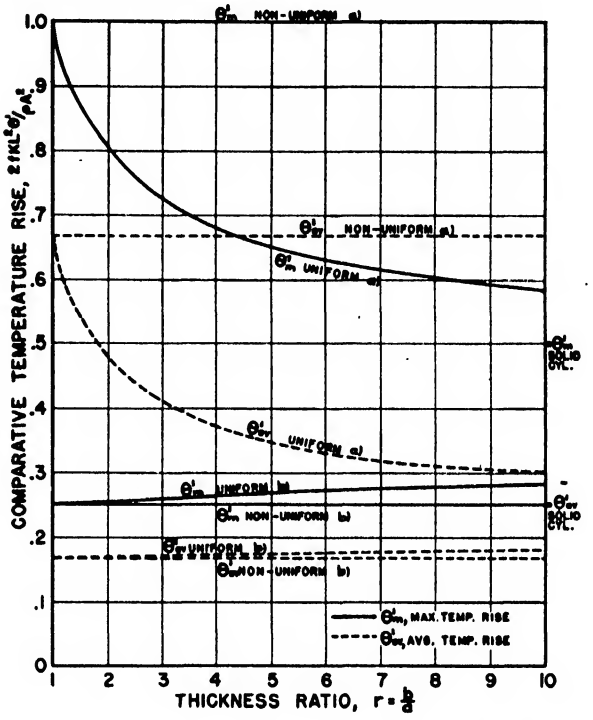


FIG. 2. Maximum and average temperature rise with constant resistivity.

of the radius of the turns in the winding. Thus the value of Q is given by the formula

$$Q = \rho A^2/[fL^2x^2(\log_e r)^2]. \quad (21)$$

Assuming small temperature rise conditions, both ρ and K are constant; hence the temperature field in a non-uniform cylindrical winding is represented by the equation

$$\frac{d^2T'}{dx^2} + \frac{1}{x}\frac{dT'}{dx} + \frac{1}{x^2}\cdot\frac{\rho A^2}{fKL^2(\log_e r)^2} = 0. \quad (22)$$

With boundary conditions (a), the temperature distribution is found to be

$$\Theta'(x) = \frac{\rho A^2}{2fKL^2(\log_e r)^2}\log_e\frac{b}{x}\log_e\frac{bx}{a^2} \quad (23)$$

[6] C. L. Emmerich, Am. Math. Monthly 55, 613 (1948).

which yields as the maximum temperature rise the value already found in Eq. (5). However, while Eq. (5) was found as the solution to the problem of the uniform winding of thickness ratio one only, the solution just determined is independent of the thickness ratio, hence representable by the straight line marked "Θ_m' non-uniform (a)" in Fig. 2.

Experimental data[7] indicate that the practical approximations to ideal non-uniform windings[8] exhibit the same temperature rise characteristics as the ideal windings. Thus the solution just determined for the ideal non-uniform windings is valid for practical work also.

The remarkable property that non-uniform windings exhibit the same temperature rise, independent of thickness ratio, is shown also by the solution with boundary conditions (b), as indicated in Fig. 2 by the straight line "Θ_m' non-uniform (b)," as well as the average temperature rises "Θ_{Av}' non-uniform (a)" and "Θ_{Av}' non-uniform (b)."

The preceding results show quantitatively to what extent overheating may be expected in multiple-wire-size windings, when thin wire is wound near the core as suggested by Emmerich.[6] The conclusion is that the temperature rise is the same as for a uniform winding of the same number of ampere-turns, but wound on a core with zero curvature.

LARGE TEMPERATURE RISE

Insulating materials which were developed recently have made increased operating temperatures possible for magnet coil windings.[9] In order to take full advantage of these new materials, it is imperative to have correct design data at hand. These data are obtainable in part from the solution of Poisson's Eq. (1).

When large temperature rises are considered, it is no longer possible to assume that Q and K are constant. The rate of heating, Q, is a linear function of the temperature, because it contains as a factor the resistivity ρ, which is known to vary linearly with temperature in the range normally expected for operating conditions of electrical equipment. The average thermal conductivity K, being a function of the conductivities of the materials in the coil, may be expected to vary with temperature also; however, it appears difficult to predict the relationship between K and T for any given winding. Experimentally, a coil of Formex wire, immersed in kerosene, was found to have a constant thermal conductivity of 0.020 watt/°C-inch over a temperature range from 12°C to 65°C; while the same coil in air exhibited an increase in K from 0.0057 to 0.0072 watt/°C-inch as the temperature changed from 33°C to 84°C.[7]

Thus, at least for a Formex winding immersed in kerosene, the right side of Eq. (1) should be expressed as a linear function of T. It will be helpful in applying the boundary conditions to let the temperature be represented by the sum of the surface temperature and temperature rise:

$$T=T_s+\Theta.$$

The differential operator ∇^2 will then apply directly to the temperature rise Θ, while the linear function in T is replaced by a linear function in Θ. Poisson's equation then reads as follows

$$\nabla^2\Theta=-u^2\Theta-w^2 \tag{24}$$

where

$$u^2=\frac{\alpha_0\rho_0A^2}{fKL^2h^2} \quad \text{and} \quad w^2=\frac{\rho_0A^2}{fKL^2h^2}(1+\alpha_0T_s).$$

In these expressions, ρ_0 is the resistivity of the metal in the wire at 0°C, and α_0 is the temperature coefficient of resistivity of the wire at 0°C. The boundary conditions are simplified to the extent that at the outside surface of the winding, the temperature rise is zero.

Jakob[5] has published the solutions of Eq. (24) for the solid cylinder ($r\to\infty$) and the flat plate ($r=1$). The maximum temperature rise in the case of the flat winding ($r=1$) is given as follows, in the notation of this paper,

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)\left[\sec\frac{A(\alpha_0\rho_0)^{\frac{1}{2}}}{L(fK)^{\frac{1}{2}}}-1\right]. \tag{25}$$

From the expansion of the secant function it is readily found that Eq. (25) is approximated by the series

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)\left[\frac{1}{2!}u^2h^2+\frac{5}{4!}(u^2h^2)^2+\frac{61}{6!}(u^2h^2)^3+\cdots\right]. \tag{26}$$

It is of course possible to use Eq. (25), or its expansion (26), directly in making temperature rise calculations. However, for the purpose of comparison with the design data obtained from Eq. (5), which neglects changes in resistivity, a correction factor Θ_m/Θ_m' has been computed.

If the value of ρ in Eq. (5) is taken as the resistivity of the wire at temperature 0°C, it will be seen that the temperature rise computed from Eq. (5) is

$$\Theta_{m0}'=u^2h^2/2\alpha_0. \tag{27}$$

Substituting this expression in Eqs. (25) and (26) leads to the functional equation

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)[\sec(2\alpha_0\Theta_{m0}')^{\frac{1}{2}}-1] \tag{28}$$

[7] C. L. Emmerich, thesis, University of Cincinnati (1949), Section V.
[8] C. L. Emmerich, Elec. Eng. 68, 185 (1949).
[9] G. L. Moses and J. J. Torok, Trans. A.I.E.E. 65, 412 (1946).

and to the series expansion

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)\left[\alpha_0\Theta_{m0}'+\frac{5}{4!}(2\alpha_0\Theta_{m0}')^2+\frac{61}{6!}(2\alpha_0\Theta_{m0}')^3+\cdots\right]. \quad (29)$$

If the value of ρ in Eq. (5) is taken as the resistivity of the wire at temperature T_s, that is, if

$$\rho_s=\rho_0(1+\alpha_0 T_s)$$

as is often done in coil design, then substitution yields the functional equation

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)\left[\sec\left(\frac{2\Theta_{ms}'}{T_s+1/\alpha_0}\right)^{\frac{1}{2}}-1\right] \quad (30)$$

and the corresponding expansion

$$\Theta_m=\Theta_{ms}'+\frac{5}{6}\cdot\frac{1}{(T_s+1/\alpha_0)}(\Theta_{ms}')^2+\frac{61}{90}\cdot\frac{1}{(T_s+1/\alpha_0)^2}(\Theta_{ms}')^3+\cdots. \quad (31)$$

Dividing Eq. (30) or (31) through by Θ_{ms}' yields the correction factor Θ_m/Θ_{ms}', which is a function of Θ_{ms}' with T_s as a parameter. This family of curves is plotted in Fig. 3, with $1/\alpha_0=234.5$°C.

An example will illustrate the use of these curves: At a surface temperature $T_s=40$°C, the uncorrected temperature rise obtained from Eq. (5) for a particular winding is 84°C, using the value of ρ_s corresponding to $T_s=40$°C. On the curve marked $T_s=40$°C in Fig. 3, the correction factor corresponding to 84°C is 1.35. Hence, the true temperature rise is $1.35\times84=113$°C. The "hot spot" in the coil will therefore reach a temperature of $40+113=153$°C.

It should be kept in mind in using these data that the boundary conditions are idealized. The assumption is made that no heat escapes through the core or through the ends of the coil, and that the temperature at the surface is constant. In actual windings it is probable that heat does escape through the core and ends, thus making the foregoing calculations fall on the safe side; however, the temperature at the surface of the coil may be quite far from the ambient temperature, and it may vary along the surface. This tends to make the calculations uncertain. To take into account the end leakage leads to rather involved computations, as shown by Higgins.[10] Actual surface temperatures are probably obtained easiest by experimental methods.[2-4]

It should be emphasized that Eq. (30) and Fig. 3 apply only to windings of thickness ratio one. Since actual windings are usually in the shape of hollow cylinders, the following computations have been made to evaluate the corrections necessary to allow for the curvature of the winding. It will be found, as in the case of small temperature rise, that the winding of thickness ratio one has a higher maximum temperature rise than a winding in the shape of a hollow cylinder.

For a long, thick, cylindrical winding (heat flow in the radial direction only), with boundary conditions (a), the solution of the equation

$$\frac{d^2\Theta}{dx^2}+\frac{1}{x}\frac{d\Theta}{dx}+u^2\Theta=-w^2 \quad (32)$$

will permit calculation of the temperature rise distribution

$$\Theta(x)=\frac{w^2}{u^2}\left[\frac{J_1(au)Y_0(xu)-J_0(xu)Y_1(au)}{J_1(au)Y_0(bu)-J_0(bu)Y_1(au)}-1\right] \quad (33)$$

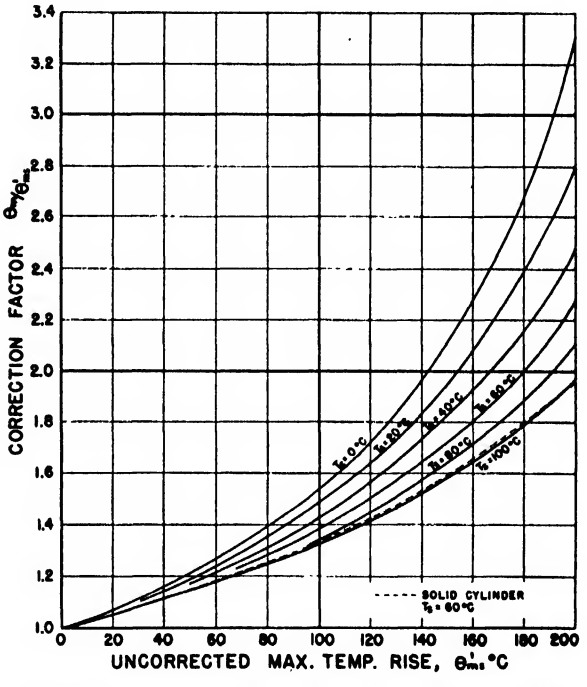


FIG. 3. Correction factor for maximum temperature rise in thin windings.

and of the maximum temperature rise

$$\Theta_m=\left(T_s+\frac{1}{\alpha_0}\right)\times\left\{\frac{2}{\pi au[J_1(au)Y_0(bu)-J_0(bu)Y_1(au)]}-1\right\}. \quad (34)$$

J and Y are Bessel functions of the first and second kind, their order being indicated by subscripts.

[10] T. J. Higgins, Trans. A.I.E.E. 64, 190 (1945) and Trans. A.S.M.E. 66, 665 (1944).

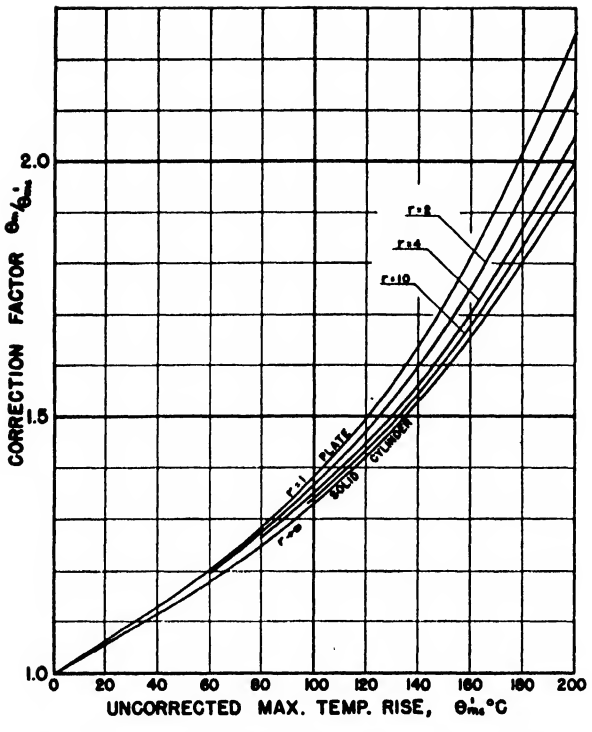


FIG. 4. Correction factor for maximum temperature rise in thick windings, $T_s = 60°C$.

Again, Eq. (34) might be used directly in making temperature rise calculations. However, for the purpose of comparison with design data obtainable from Eq. (14), a correction factor Θ_m/Θ_{ms}' has been computed and plotted in Fig. 4, for the surface temperature $T_s = 60°C$, and using the value of the reciprocal of the thermal coefficient of resistance $1/\alpha_0 = 234.5°C$.

This family of curves is intended to demonstrate that the difference in correction factors for windings of various thickness ratios, other conditions being the same, is rather small. Few actual windings exceed the thickness ratio $r = 2$; thus the approximation made by using the curves in Fig. 3 is not serious. For the purpose of comparison, the curve for a solid cylindrical winding at $T_s = 60°C$ has been shown dotted in Fig. 3.

It may be stated here for the sake of completeness that the non-uniform winding designed for minimum resistance at small temperature rise yields exactly the same temperature rise conditions as the curves in Fig. 3, independent of its thickness ratio.

Up to this point, Poisson's Eq. (1) has been solved for cylindrical windings with Q and K constant; also with K constant and Q a linear function $Q(T)$. If it is assumed that K is also a linear function $K(T)$, then the differential equations become much more complicated, making a solution in closed form difficult. The only exception occurs when $K(T)$ is a multiple of $Q(T)$; that is, when the linear function $K(T)$ would be zero at the same temperature as the temperature for which $Q(T)$ would be zero. Then the right side of Eq. (1) would reduce to a constant again. In practical terms, this requires that the thermal conductivity of the winding must be proportional to the resistivity of the metal in the wire at all temperatures. Ordinary coil windings are not expected to have this special property.

If the function $1/K$ is linear in T, then the right side of Eq. (1), being the product of two linear functions, would become a quadratic function in T. This case has already been considered by Davies,[11] who assumed K to be constant, and $Q(T)$ a quadratic function. The solution in this case contains elliptic functions. It appears that the complete and detailed results of this case are not very useful until it has been established that the thermal conductivities of actual windings are expressed correctly by linear functions $1/K$.

ACKNOWLEDGMENT

Grateful acknowledgment is made to the Stephen H. Wilder Foundation, and to the staff of the Applied Science Research Laboratory, University of Cincinnati, for their cooperation on this project.

[11] E. S. Davies, Trans. A.S.M.E. 65, 602 (1943).

Evaporation and Outgassing in an Inert Atmosphere

R. L. LONGINI
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
(Received June 9, 1949)

Experiments on the evaporation outgassing of antimony in vacuum and in an inert atmosphere show that an inert atmosphere is advantageous for outgassing antimony if loss of antimony by evaporation is intollerable. A simple theory based on well-established kenetic theory considerations is given which may explain and did predict the results.

INTRODUCTION

THE introduction of an inert gas into a vacuum system changes the conditions limiting the rates of both evaporation and outgassing. In a vacuum these rates depend on reflection or accommodation coefficients, whereas in the presence of an inert atmosphere diffusion is the rate-determining process.

It has been known for some time that an atmosphere reduces the rate of evaporation of tungsten in lamps or of magnesium[1] in an inert atmosphere.

In the present study, the actual rates were determined experimentally by use of a microbalance on which loss of sorbed gasses and of the parent body of antimony were found by direct weighing of test samples. A helium atmosphere was used. Results indicate that the introduction of an inert atmosphere may prove useful in outgassing volatile materials.

THEORY

The theory is rather crude and may neglect some important effects. It is important to note that the theory came first and predicted the observed results. It is presented as it may prove useful in some other cases.[2]

Let us assume that the test sample is in equilibrium with its various vapors. The vapor of the parent material (antimony in the experiments) is in equilibrium with the solid body, and other vapors are in equilibrium with their sorbed phases.

Suppose we examine any one of the vapors closely. Let n_{i0} be the number of molecules per unit volume of the ith vapor in the space at equilibrium. Now kinetic theory shows us that $n_{i0}\bar{c}_i/4$ of these will cross a unit area per second where $\bar{c}_i$ is the mean velocity of the ith vapor.[3] This will also be the number of molecules striking each unit area of the condensed material, and a fraction a_i will stick. The fraction a_i is known as the accommodation coefficient. At equilibrium, the rate of condensation or sticking must equal the gross rate of evaporation of the ith component R_i, which gives

$$R_i = n_{i0}\bar{c}_i a_i/4. \tag{1}$$

(If a good vacuum is maintained the net rate of evaporation must be given by R_i above, as the vapor will be removed and no condensation from a vapor phase will take place.)

Suppose we admit an inert atmosphere. The evaporating vapors will now have to diffuse away from the condensed phase. Let R_{pi} be this evaporation rate when the inert gas pressure is p. Now $R_{pi}=D_{pi}(\text{grad}n_i)_0$ where D_{pi} is the diffusion coefficient of the ith vapor through the inert atmosphere of pressure p and n_i is the number of molecules of the ith vapor per cubic centimeter. The gradient is taken at the condensing surface as indicated by the subscript zero.

Now, from our previous discussion, we conclude that the vacuum or gross rate R_i is equal to the net rate R_{pi} plus the rate of recondensing $n_{i0}a_i\bar{c}_i/4$, where n_{i0} is the molecular density at the surface.

We can state

$$\frac{R_{pi}}{R_i}=\frac{D_{pi}(\text{grad}n_i)_0}{n_{i0}\bar{c}_i a_i/4+D_{pi}(\text{grad}n_i)_0}. \tag{2}$$

From kinetic theory we have $D_{pi}=g_i T^{\frac{1}{2}}p^{-1}$ and $\bar{c}_i=h_i T^{\frac{1}{2}}$, where g_i and h_i are proportionality constants. Putting these in the above and dividing numerator and denominator each by $g_i(\text{grad}n_i)_0 T^{\frac{1}{2}}$ we have

$$\frac{R_{pi}}{R_i}=\frac{T/p}{T/p+a_i n_{i0}h_i/4g_i(\text{grad}n_i)_0}. \tag{3}$$

In the steady state it can be shown that $n_{i0}/(\text{grad}n_i)_0$ is independent of n_i and is a function of geometry only. (In the one-dimensional case, for example, $n_{i0}/(\text{grad}n_i)_0$ equals the distance between the evaporating surface and the remote absorbing surface.) It is therefore essentially the same for all components. We can rewrite the above as

$$R_{pi}=R_i\frac{T/p}{a_i/G_i+T/p}, \tag{4}$$

where G_i takes into account both the geometry and the proportionality constants for the ith vapor. Now the diffusion coefficients for the various vapors are of the same order of magnitude so that the G_i and thus the ratios R_{pi}/R_i of various components have large differences only because of large differences in the accommodation coefficients.

[1] E. A. Gulbransen, Trans. Electrochem. Soc. 87, 597 (1945).
[2] For comparison see I. Langmuir, Phys. Rev. 2, 329 (1913).
[3] For example, *Cobine, Gaseous Conductors* (McGraw-Hill Book Company, Inc., New York, 1941), p. 19.

The accommodation coefficient for the parent material will probably be close to unity, whereas that for various sorbed vapors may be very small. Thus a pressure of inert gas may do much more to reduce the net evaporation for the parent material than for the sorbed components. The ratio of rates of evaporation of two components is thus pressure dependent

$$\frac{R_{p1}}{R_{p2}} = \frac{R_1}{R_2}\,\frac{T/p+a_2/G_2}{T/p+a_1/G_1}. \tag{5}$$

Anticipating our experimental results with plated antimony and its various sorbed vapors,

$$\frac{R_{p(Sb)}}{R_{p2}} = \frac{R_{(Sb)}}{R_2}\,\frac{T/p+3.3\times10^4}{T/p+4.1\times10^6} \tag{6}$$

where the subscript Sb refers to plated antimony, the subscript 2 refers to an effective "mean" of sorbed vapors. The pressure is in atmospheres, the temperature in degrees Kelvin. In this case the value of $n_{i0}/(\text{grad}n_i)_0$ was about 25 cm. It is apparent that a moderate pressure will make the T/p value small and that the rate of evaporation $R_{p(Sb)}$ will be reduced far more than the rate of outgassing R_{p2}.

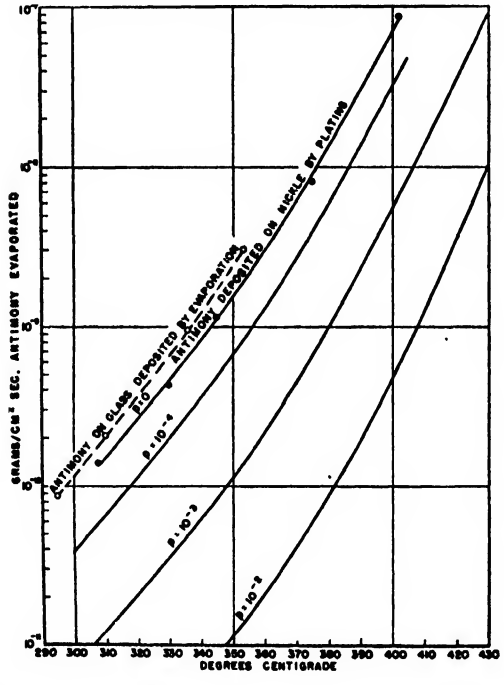


FIG. 1. The rate of evaporation of plated antimony. The rate of evaporation of plated antimony is shown as a function of temperature and helium pressure. The pressures of the isobaric curves are expressed in atmospheres. The vacuum rate of evaporation of antimony from glass is also shown. The experimental points for the vacuum rates are given but those for inert atmospheres were very scattered, varying by a factor of two from the curves shown. Since this is influenced by geometry in any case, the curves are given only to show the order of magnitude of the effect.

EXPERIMENTAL PROCEDURE AND RESULTS

The experiments to determine evaporation rates were carried out on Dr. E. A. Gulbransen's microbalance.[4] This instrument is enclosed in glass and can be evacuated or filled to pressures up to atmospheric with nearly any gas. The sample is suspended from one side of the balance by a tungsten wire so positioned that the sample hangs in a tube which can be heated by a furnace. The counterweight has the same density as the sample to reduce buoyancy effects when gas is admitted.

The balance is calibrated in micrograms and reads the degree of unbalance between the constant counterweight and the sample under investigation. The readings are made with a vernier microscope and are reproducible under most conditions to within one-half a microgram.

Two types of samples were used. The first consisted of 0.005-in. steel shims 1×6 centimeters in extent, plated on both sides with a total of 16 milligrams of antimony from an antimony chloride plating bath. This type was used to determine the rate of evaporation of outgassed plated antimony, the rate of outgassing, and the effect of helium on these rates. The second type consisted of microscope cover glasses each 11×50 millimeters by 0.25 millimeter thick. These were coated with antimony which was deposited by evaporating it from a tungsten wire on which it had been plated. The resultant coating had a surface density of about 16 μg cm^{-2} on each side. These slides were used for determining the rate of evaporation from an evaporated antimony film on glass.

For determining the rate of evaporation of antimony the samples were first outgassed. When the rate of weight loss was constant with time at any one temperature, the sample was believed to be outgassed. This was cross-checked by raising the temperature for a while, then bringing it back to the original temperature. A return of the rate of weight loss to its original steady value was always found.

The various rates were determined by plotting the loss of weight against time at a fixed temperature and observing the slope. During outgassing the resulting plots were curves with steadily, but slowly, decreasing slopes. The slopes of the steady state lines corresponding to outgassed samples were found and, after being put in terms of grams cm^{-2} sec.$^{-1}$, are plotted against temperature in Fig. 1. The rates of evaporation of both forms of antimony have the same slope when plotted against temperature but the evaporated film has a 40 to 50 percent higher rate. This is indicative of the same heat of vaporization but probably of a higher accommodation coefficient for the evaporated film. This indicates an accommodation coefficient of less than 0.7 for plated antimony, a very low value compared to the assumed value of unity used in calculating the calibra-

[4] E. A. Gulbransen, Rev. Sci. Inst. 15, 201 (1944).

tion of the Knudsen gauge. The apparent higher accommodation coefficient for the evaporated film is probably due to a granular structure similar to those seen with an electron microscope for most metal films. The condensing antimony molecule, falling deep into the layer, might have to undergo several reflections before emerging.

It is interesting to note that the evaporation rate of antimony from the glass slide was independent of the amount of antimony on the slide until this was under 2 $\mu g\ cm^{-2}$. Below this the rate dropped slowly to one-half its former value, then dropped suddenly to zero at which point all the antimony had evaporated.

Outgassed samples were observed when a helium atmosphere was introduced. It was found that even slight impurities in the helium had a very decided effect on the results. The helium was passed through hot, partially oxidized copper and then through a drying tube. This should have removed hydrogen, oxygen, and water vapor. Even after great care with this process, the rates determine with helium present were not reproducible with precision. The determined effect of pressure may be off by 50 percent. The figure shows how the rate of evaporation is dependent on temperature and pressure.

Determining the effect of an inert gas pressure on the rate of outgassing was much more difficult. First of all, the outgassing rate in vacuum is not constant so that the pressure effect had to be measured rapidly. Secondly, the effect was so small that 0.1 atmosphere of helium was needed to measure it. These conditions required a much more rapid introduction of gas than the previously described purifier could handle. A long total obstruction activated charcoal trap was built to purify the gas more rapidly.

The tests were carried out as quickly as possible. A rate was determined for the vacuum condition, (1.5×10^{-3} $\mu g\ sec.^{-1}\ cm^{-2}$ at 284°C), then helium at a pressure of 0.1 atmosphere was introduced and a new rate determined; (2.17×10^{-4} $\mu g\ sec.^{-1}\ cm^{-2}$) the helium was then pumped out and the vacuum rate determined again and was found unchanged. A slight buoyancy effect, probably caused by convection currents in the helium, caused jumps in the readings on introduction and exhaust of the gas. The jumps were equal and opposite so that the measured evaporation rate with a gas atmosphere was assumed to be correct. The outgassing rate remained constant during the run with 0.1 atmosphere of helium present. The total loss of weight during the test was about 50 μg. The sample was of antimony plated on steel shim stock. The test was carried out at 284°C.

As Fig. 1 shows the evaporation rate of antimony in vacuum at 284°C was approximately 6×10^{-5} $g/cm^2/sec.$ or one twenty-fifth the rate of outgassing in vacuum at this temperature. During the inert atmosphere part of the test at 0.1 atmos. the outgassing rate was reduced by a factor of 7, but, as Fig. 1 shows, the evaporation rate was reduced by a factor of 100 to 1000. Thus, the introduction of the inert gas had the effect of reducing the evaporation rate to a very small fraction of the outgassing rate.

CONCLUSION

An inert gas requires that evaporated molecules diffuse through it and therefore it limits the rate of evaporation. Many evaporated molecules are reflected by the inert gas back to the evaporating surface where they may recondense.

A difference in the accommodation coefficient of this condensing process may explain completely the alteration of the relative rates of outgassing and evaporation of the parent body when an inert atmosphere is introduced. The introduction of such an atmosphere may prove useful when it is desired to outgas other volatile material, for the evaporation rate is probably reduced far more than the outgassing rate in most cases. In the particular case of plated antimony, for the same rate of outgassing, the rate of loss of antimony by evaporation in an inert atmosphere was less than one one-hundredth the rate of loss in a vacuum.[5]

[5] U. S. Patent No. 2,456,968.

A Method of Reducing Deflection Defocusing in Cathode-Ray Tubes

R. G. E. Hutter and S. W. Harrison
Physics Laboratories, Sylvania Electric Products, Inc., Bayside, New York
(Received May 26, 1949)

A discussion of deflection defocusing is given to point out the two-dimensional character of the focusing action of deflection fields. A photograph of spot distortion actually observed is shown. A brief theory of combined spherical and cylindrical lenses is given, and a method of spot correction based on the principle of "beam predistortion" is described. An electron gun used for such correction is shown and experimental results obtained with this gun are exhibited.

INTRODUCTION

BOTH practical experience and theoretical analysis[1–3] show than an electron beam passing through a deflection field is subject to a "two-dimensional" focusing action as well as a bending action. In cathode-ray tubes, such a focusing action is undesirable. If the beam is focused to a small spot at the center of the fluorescent screen and is then deflected, the result of the two-dimensional focusing action will be a distortion of the spot.*

The amount of beam distortion produced by a deflection field varies with the form of the deflection field employed. Attempts to reduce the distortion effects by improving the design of the fields have been relatively successful in the case of magnetic deflection. This has not been true in the case of electrostatic fields where one is more limited in the type of change that can be made in the elements producing the field. The fact that there is a change in the velocity of the electrons as well as a variation in the strength of the electrostatic field also makes defocusing effects in this type of field more inherent than in the magnetic field where there is only a variation in field strength.

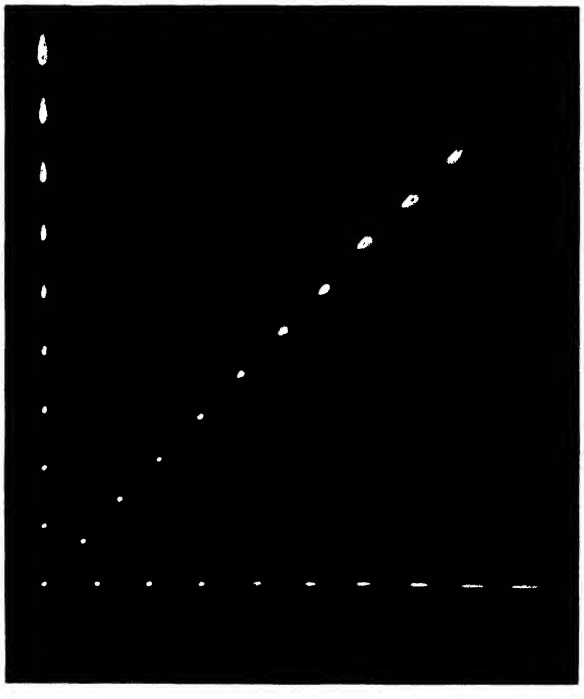


FIG. 1. Spot distortion observed with an electrostatic deflection system of low sensitivity.

The amount of beam distortion associated with any deflection depends also on the parameters of the beam as it enters the field, For a given deflection, a certain set of initial conditions will give a beam which is free of distortion after deflection. Physically, this may be interpreted as saying that the beam is initially distorted in a manner such as to compensate for the distortion produced by the deflection field. To produce such a predistortion, the electron optics of an ordinary axially symmetric focusing system must be modified. The magnitude of this modification will depend on the magnitude and direction of deflection, or, in other words, on the position of the spot on the fluorescent screen.

It is the purpose of this paper to describe a method of spot correction which is based on the principle of beam predistortion. This principle can be employed in both the electrostatic case and the magnetic case, but the particular method described is electrostatic in nature and is used in connection with electrostatic deflection systems.

In describing the method of spot correction it is convenient to begin with a brief discussion of deflection fields. It will be pointed out that the beam distorting action which is associated with deflection is "two-dimensional" in nature. Observed spots will be shown which confirm the character of this type of distortion. This discussion of deflection fields and the effect of such fields on electron beams will be followed by a first-order theory of focusing fields which have spherical and cylindrical lens properties. It will then become apparent that, by combining such focusing fields with deflection fields, the spot distortion effects can be reduced.

Finally, a special electron gun will be described. This

[1] J. Picht and J. Himpan, Ann. d. Physik 39, 409 (1941); 43, 53 (1943).
[2] G. Wendt, Die Telefunkenrohre 15, 100 (1939); Zeits. f. Physik 118, 593 (1942).
[3] R. G. E. Hutter, J. App. Phys. 18, 740–758 (1947).
* The properties of a deflection field also cause a non-linear relationship to exist between field strength and deflection. This in turn produces a so-called "pattern distortion." Since this is ordinarily less objectionable, it will be ignored in the present discussion.

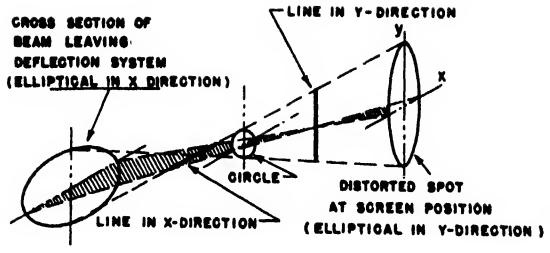


FIG. 2. Distortion which results when an electron beam, point focused before deflection, is deflected in the y-direction.

gun is capable of producing an undistorted spot at any point on the screen to which the beam can be deflected. This means that, at the fluorescent screen, the cross section of a deflected beam is circular and has nearly the same diameter as the cross section of the undeflected beam. Photographs taken of the face of a special tube in which this gun was incorporated will show the degree of spot correction achieved.

I. BEAM DISTORTING EFFECTS OF DEFLECTION FIELDS

Actual cathode-ray tubes employing balanced electrostatic deflection fields are designed to have as little spot distortion as possible by keeping the angle of deflection small. In order to study methods of correcting the spot distortion, an electron gun of low sensitivity and hence large distortion effects was designed. A photograph of spots obtained with this gun is shown in Fig. 1 and it can be seen that the general character of the spot distortion is as described by the theory.[1-3]

Both the calculations and the observed results show that the spot is elongated in the direction of deflection. A closer examination of the calculated results, as well as a closer analysis of experimental results for an initially point-focused beam, shows that this characteristic of the spot distortion is due to a focusing effect of the deflection field which is considerably stronger in the direction of deflection than perpendicular to it. The shape of the deflected electron beam on the gun side of the screen is sketched in Fig. 2.

It can be seen from this that in no portion of the beam is there a convergence of all the electron paths

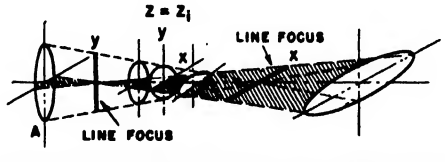


FIG. 3. Shape of an electron beam in the neighborhood of the image plane of a combined spherical and cylindrical electron lens.

to a point. At two places the cross section degenerates into a straight line and at an intermediate point a circular cross section can be found. All other portions of the beam have an elliptical cross section. Reducing the strength of the focusing field of the electron gun will make all paths less convergent, and it is possible to adjust this strength so that the electron beam has a circular cross section at the screen. However, no point focus is now possible in contrast with the point focus assumed at the center of the screen. In the practical case the spot at the center has a finite radius, but no refocusing of the electron lens of rotational symmetry will produce a deflected spot of the same size as the undeflected spot.

A cylindrical lens action must be incorporated in the electron gun to obtain perfect correction for both shape and size. In order to show that this is the case, a brief discussion will be given of the theory of a class of focusing fields which have both spherical and cylindrical lens properties.

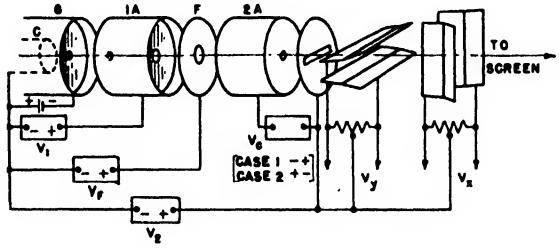


FIG. 4. Electrode arrangement and associated power supplies for an electron gun free of deflection defocusing.

II. THEORY OF COMBINED SPHERICAL AND CYLINDRICAL ELECTRON LENSES

The potential distribution of an electrostatic field which satisfies the symmetry condition,

$$\varphi(x,y,z)=\varphi(-x,y,z)=\varphi(x,-y,z) \tag{1}$$

may be expanded in the series:

$$\varphi(r,z,\theta)=\phi(z)-(1/4)\phi''(z)r^2+\phi_2(z)r^2\cos 2\theta +(1/64)\phi^{IV}(z)r^4-(1/12)\phi_2''(z)r^4\cos 2\theta +\phi_4(z)r^4\cos 4\theta+\cdots. \tag{2}$$

For a field of rotational symmetry $\phi_{2n}(z)=0$. The paraxial ray equations for fields given by Eq. (2) are

$$\left.\begin{aligned} x''+(1/2)(\phi'/\phi)x'+(1/4)(\phi''/\phi)x&=(\phi_2/\phi)x \\ y''+(1/2)(\phi'/\phi)y'+(1/4)(\phi''/\phi)y&=-(\phi_2/\phi)y \end{aligned}\right\}. \tag{3}$$

These equations may be integrated under the following conditions:

(1) The general solution of the paraxial ray equation is known for the case $\phi_{2n}(z)=0$ (spherical lens).

(2) Terms which contain $\phi_{2n}(z)$ (combined spherical and cylindrical lens) may be considered small distortion terms.

With these conditions, the integration of Eq. (3) may be carried out by employing the methods ordinarily used in lens aberration theory.

Let $r(z)$ be the general solution of the paraxial ray equation:

$$r''+(1/2)(\phi'/\phi)r'+(1/4)(\phi''/\phi)r=0. \tag{4}$$

If $r(z)$ has the value r_0 in the object plane $z=z_0$ and

the value r_a in the aperture plane $z=z_a$, the solution may be written

$$r(z)=r_0 r_\gamma(z)+(r_a/r_{aa})r_\alpha(z), \tag{5}$$

where $r_\alpha(z)$ and $r_\gamma(z)$ are two particular solutions which

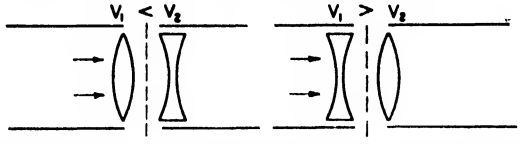


FIG. 5. A simple two-cylinder rotationally symmetric electron lens and the optical analogies for $V_1<V_2$ and $V_1>V_2$.

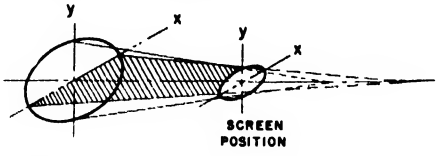


FIG. 6. Distortion of an undeflected electron beam necessary to compensate for the focusing action associated with a deflection in the x-direction.

satisfy the conditions

$$\begin{aligned} r_\alpha(z_0)&=0 \quad r_\alpha'(z_0)=1 \\ r_\gamma(z_0)&=1 \quad r_\gamma(z_a)=0. \end{aligned} \tag{6}$$

The values of $x(z)$ and $y(z)$ of Eq. (3) in the image plane are then given** by:

$$\left.\begin{aligned} x(z_i)=\Delta x_i&=-\frac{M}{(\phi_0)^{\frac{1}{4}}}\int_{z_0}^{z_i}\frac{\phi_2(z)}{(\phi(z))^{\frac{1}{4}}}r(z)r_\alpha(z)dz \\ y(z_i)=\Delta y_i&=+\frac{M}{(\phi_0)^{\frac{1}{4}}}\int_{z_0}^{z_i}\frac{\phi_2(z)}{(\phi(z))^{\frac{1}{4}}}r(z)r_\alpha(z)dz \end{aligned}\right\}. \tag{7}$$

For electron paths for which $r_0=0$, Eqs. (5) and (7) may be combined to give

$$\left.\begin{aligned} \Delta x_i&=-\frac{M}{(\phi_0)^{\frac{1}{4}}}\frac{r_a}{r_{aa}}\int_{z_0}^{z_i}\frac{\phi_2(z)}{(\phi(z))^{\frac{1}{4}}}r_\alpha^2(z)dz \\ \Delta y_i&=+\frac{M}{(\phi_0)^{\frac{1}{4}}}\frac{r_a}{r_{aa}}\int_{z_0}^{z_i}\frac{\phi_2(z)}{(\phi(z))^{\frac{1}{4}}}r_\alpha^2(z)dz \end{aligned}\right\}, \tag{8}$$

where $\phi_0=\phi(z_0)$ and $M=r_\gamma(z_i)$ is the magnification of the rotationally symmetric lens. Equation (8) shows that $\Delta x_i=-\Delta y_i$ and the cross section of the electron beam in the neighborhood of the image plane will vary as is shown in Fig. 3.

The type of lens described above may be used to produce a beam which converges differently in the x- and y-direction. This is the effect which was observed in the defocusing action of a deflection field. Such a lens may thus be used to compensate for the asymmetrical focusing action of a deflection system.

** See Madelung, *Die Mathematischen Hilfsmittel des Physikers* (Dover Publications, New York, 1943), third edition, p. 183.

III. AN ELECTRON GUN FREE OF DEFLECTION DEFOCUSING

On the basis of the analysis given above, it is possible to design an electron gun which is free of deflection defocusing. Such a gun is shown schematically in Fig. 4.

With the exception of a slit electrode, this gun is similar to the electron gun used in Type 7GP4 cathode-ray tubes. The slit is parallel to the first pair of deflection plates. The first and second anode ($1A$ and $2A$) are usually operated at the same high potential with respect to the cathode (i.e., $V_1=V_2$, $V_c=0$). The focusing electrode (F) is operated at a fraction of the value of the anode potential and the voltage of F is adjusted to produce as small a spot as possible on the screen. Because of space charge effects, velocity distribution effects, and lens aberrations, this spot is not a true point focus. The deflection systems are balanced as indicated in Fig. 4. The space between the last aperture in the second anode and the screen is at the uniform potential of the last anode if the deflection potentials V_x and V_y are equal to zero.

The addition of the slit electrode will not change anything when $V_c=0$ and $V_1=V_2$ provided the slit does not intercept electrons and does not distort the field in the neighborhood of the last aperture. If, however, $V_c\neq 0$, the field between an electrode of circular symmetry ($2A$) and one of two-dimensional symmetry (the slit electrode) will constitute a lens similar to that discussed in Section II of this paper.

In order to understand the performance of this modified gun in connection with spot correction, it should first be noted that the sign of the correction voltage V_c may be reversed. It can be reasoned that this reverses the sign of $\phi_2(z)$ in Eqs. (3) and (8). Now when $\phi_2(z)$ is positive, Δx_i is negative and Δy_i positive,

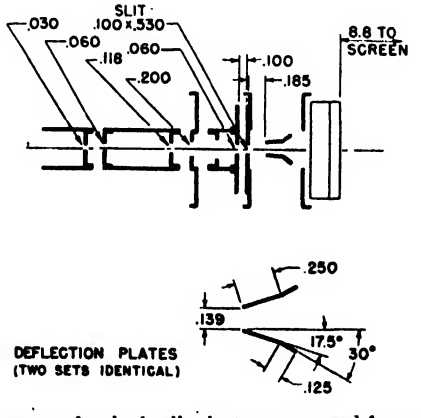


FIG. 7. Diagram of a single-slit electron gun used for correction of spot distortion. The essential dimensions in inches are shown.

and, as can be seen from Fig. 3, the beam is more convergent in the x-direction than in the y-direction. Similarly, with $\phi_2(z)$ negative, the convergence is greater in the y-direction. This reasoning presupposes a knowledge of the field distribution; since the field is

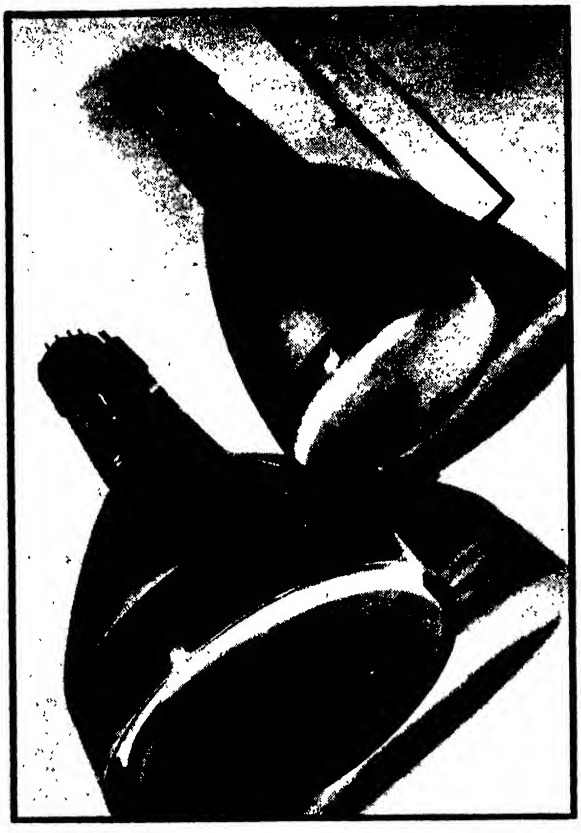


FIG. 8. Special 10-inch tube with single-slit gun and standard 7GP4 electrostatic tube showing over-all length the same for both tubes.

rather complex it is easier to proceed with a light-optical analogy.

Such analogies are often used in electron optics. Take, for instance, a simple rotationally symmetric lens which exists between two cylinders at different potentials. Such a lens is shown in Fig. 5. Although the net effect of this lens is a converging action, it should be pointed out that part of the lens has a diverging action and part a converging action. The regions in which these actions exist will depend on the potentials on the two cylinders and the two possible cases are shown, with the light optical analogies, in Fig. 5. The same thing may be said of apertured disks which separate two regions of uniform potential; this analogy holds if one of the rotationally symmetric electrodes is replaced by a slit electrode. In the latter case, of course, the lens action in the region of the slit electrode will be two-dimensional.

The performance of the modified gun may now be explained. Consider first a deflection in the y-direction and the correction for the spot distortion associated with such a deflection. With $V_c=0$, V_1, V_2 and V_f are adjusted so that the beam is "point-focused" at the center of the screen. (V_1 may or may not be equal to V_2). If the polarity of V_c is then as shown for case I in Fig. 4, an increase of V_c will mean:

(1) A reduction in the strength of the rotationally symmetric lens associated with $1A$, F, and $2A$;

(2) The creation of a "two-dimensional" diverging lens near the slit electrode and a converging, rotationally symmetric lens near the last aperture of $2A$.

With the possibility of varying V_f which changes the strength of the spherical lens, and V_c, which changes the relative convergence in the x- and y-directions, it is possible to distort the beam in such a way that its cross section at the screen position is similar to that shown at A in Fig. 3. It follows that the beam converges less in the y-direction than in the x-direction and that the rays in both the yz- and xz-planes intersect the screen before they cross the z axis. Since a given predistortion of this nature is necessary to compensate for the focusing action associated with a deflection in the y-direction, it is possible to adjust the spot at the center of the screen in such a way that the beam will be point-focused at the screen when deflected. In practice, the procedure is to deflect the beam, observe the spot distortion, and then adjust V_f and V_c so that the spot distortion is eliminated.

Since the electron gun also provides a deflection in the x-direction, it is necessary to provide for spot correction in this direction too. For this, another slit electrode following the first slit electrode with the slit rotated through 90° can be used with nearly the same potential arrangement. However, it is possible to use the single slit electrode shown in Fig. 4 for correction in both the x- and y-directions as will be shown. The polarity of V_c is chosen as illustrated for case II in Fig. 4. Now an increase of V_c will mean:

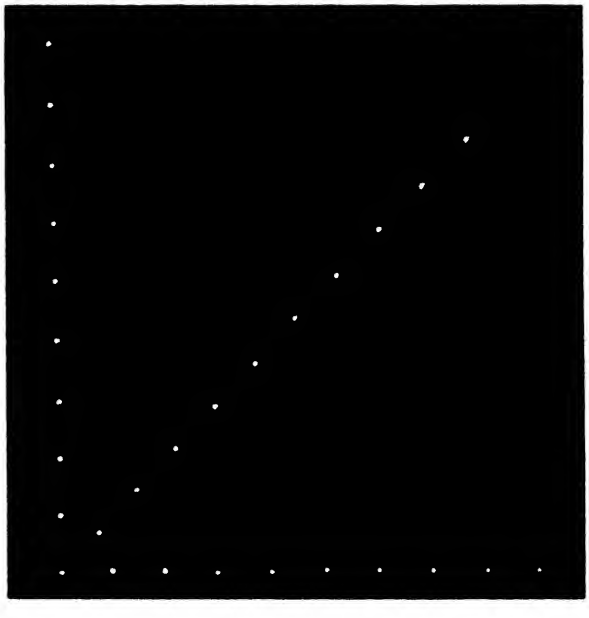


FIG. 9. Corrected spots on special 10-inch electrostatic tube with proper predistortion of the beam. The spots before correction were shown in Fig. 1.

(1) An increase in the strength of the rotationally symmetric lens associated with $1A$, F, and $2A$;

(2) The creation of a "two-dimensional" converging lens near the slit electrode and a diverging rotationally symmetric lens near the last aperture of $2A$.

Again, the possibility of varying V_f and V_c means that the beam can be predistorted, but in this case the two-dimensional lens is converging in the y-direction and the resulting beam will be less convergent in the x-direction than in the y-direction as illustrated in Fig. 6. This is the type of predistortion necessary to compensate for the focusing action associated with a deflection in the x-direction, and again the spot distortion of a deflected beam can be eliminated by adjusting V_f and V_c.

Since the spot in a cathode-ray tube is brought to an arbitrary position on the screen by deflection in both the x- and y-directions, there is some focusing in each direction due to the deflection fields. Equal amounts of focusing in each direction may be compensated for by weakening the axially symmetric focus, V_f. If the focusing action is stronger in one direction than in the other, the excess is treated just as a focusing action in one direction. Physically, this means that, for any position of the spot on the screen, it is always possible to find a combination of V_c and ΔV_f which produces a corrected spot.

IV. EXPERIMENTAL RESULTS

In an effort to check the analysis experimentally, a gun was constructed with the dimensions shown in Fig. 7. The deflection system of this gun was purposely made poor from the point of view of spot distortion by shortening the deflection plates. The gun was sealed in a ten-inch bulb with an over-all length equal to that of a standard Type 7GP4 tube. The special ten-inch tube and the 7GP4 are shown in Fig. 8. The spot distortion which results when the special tube is operated with $V_c=0$ was shown in Fig. 1. The results obtained after proper predistortion of the beam are shown in Fig. 9. The functional relationship between V_c, ΔV_f, and the deflection voltages are plotted in Figs. 10 and 11. Voltages must be applied to both pairs of deflection plates to produce a deflection along the 45° radius and, in this case, the "deflection voltage" read from the abscissa in Fig. 11 is the square root of the sum of the squares of the x- and y-deflection voltages.

It is apparent from Fig. 9 that nearly perfect spots are obtained when the predistortion is applied, even for deflections which normally produce a considerable amount of distortion. Commercial tubes have, of course, much less deflection defocusing than is obtained with the special tube without correction. However, if the angle of deflection in a regular tube were

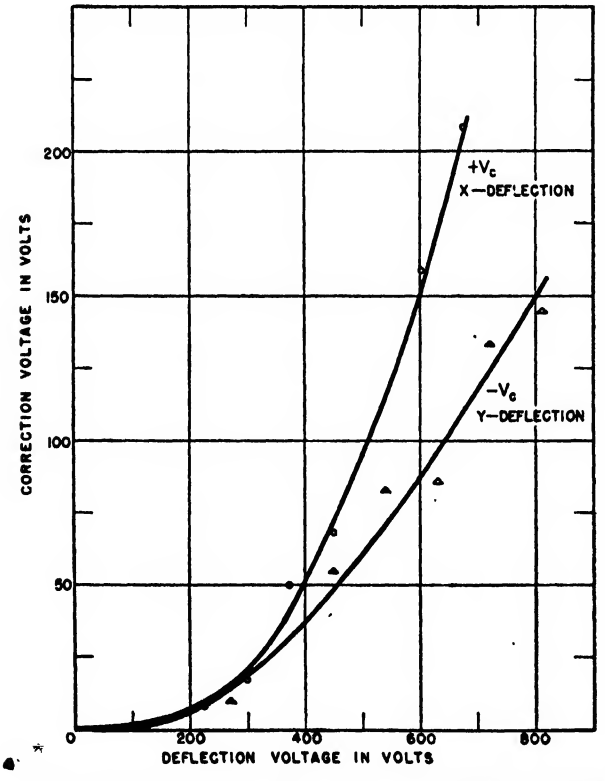


FIG. 10. Variation of correction voltages with deflection voltage for correction of spot distortion as shown in Fig. 9. The last anode voltage was 2000 volts and the grid bias (−162) volts with a pulse of 20 μsec., 110 cycles.

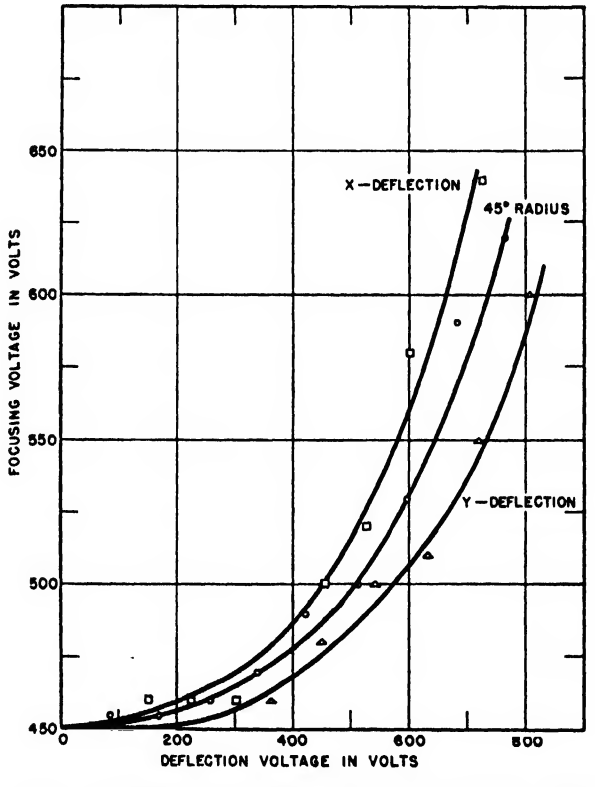


FIG. 11. Variation of focusing voltage with deflection voltage for correction of spot distortion as shown in Fig. 9. The last anode voltage was 2000 volts and the grid bias (−162) volts with a pulse of 20 μsec., 110 cycles.

increased in order to reduce the length of the tube, the spot distortion would be quite objectionable. In such a situation correction by predistortion would be very effective.

C. CONCLUDING REMARKS

The discussion above has been limited to the case of a tube employing a static deflection. In the operation of a cathode-ray tube under normal circumstances, circuits which permit automatic correction during the scanning cycle are required but will not be discussed here. The magnitudes of the correction potentials and their functional relationships to the deflection potentials depend on the particular design of the focusing and deflection systems and upon the degree of spot correction desired.

Although the correction method discussed here is for an electrostatic focusing and deflection system, it is possible to apply the principle of beam predistortion to magnetic focusing and deflection systems as well. In this connection, however, it should be pointed out that magnetic lenses, either spherical or cylindrical, are everywhere converging, i.e., there is no diverging region. If spot correction is attempted by the use of "two-dimensional" magnetic lenses, two such lenses rotated 90° with respect to each other are required.

ACKNOWLEDGMENT

The authors wish to take this opportunity of acknowledging the valuable contributions made by L. H. McKee of these laboratories in carrying out most of the experimental part of the work described.

Transmission of Elastic Waves through a Stratified Solid Medium

WILLIAM T. THOMSON*
University of Wisconsin, Madison, Wisconsin
(Received June 14, 1949)

The transmission of a plane elastic wave at oblique incidence through a stratified solid medium consisting of any number of parallel plates of different material and thickness is studied theoretically. The matrix method is used to systematize the analysis and to present the equations in a form suitable for computation.

INTRODUCTION

THE transmission of elastic waves in a homogeneous solid has been extensively treated by G. G. Stokes.[1] Barring surface waves of the type discussed by Rayleigh,[2] such waves are propagated as dilatation and rotation waves (frequently referred to as shear waves), each traveling with a different speed. It is thus necessary in dealing with such problems first to express the particle displacements in terms of the dilatation and rotation.[3]

The transmission of plane elastic waves through a stratified medium consisting of alternate parallel planes of solid and liquid layers has been treated by R. B. Lindsay[4] and others.[5,6] Since ideal fluids are incapable of sustaining tangential forces, the shearing stresses in the solid-liquid strata must vanish at the surfaces of the solid plates, thus reducing the problem to that of a single plate.

When the stratified medium consists of parallel solid plates without liquid layers between them, the problem cannot be reduced to that of a single section. The equations for one section must now be related to those of the adjoining section by the continuity of particle velocity, normal and shearing stresses at the interface. This is most conveniently accomplished by matrices which avoids unwieldly mathematical expressions and facilitates the computational task. The direction of propagation through each strata is established by Snell's law.

Nomenclature

ξ, η, ζ = particle displacement in x, y, z directions
$\dot{\xi}, \dot{\eta}, \dot{\zeta}$ = particle velocity in x, y, z directions
Δ = dilatation
ω = rotation (subscript indicates direction of normal to rotation plane)
$c_\Delta = [(\lambda+2G)/\rho]^{\frac{1}{2}}$ = velocity of dilatational wave in solid
$c_\omega = (G/\rho)^{\frac{1}{2}}$ = velocity of rotation or shear wave in solid
c_0 = velocity of propagation in the fluid
$\lambda = \nu E/[(1+\nu)(1-2\nu)]$
E = Young's modulus of elasticity
ν = Poisson's ratio
G = modulus of elasticity in shear

* Associate Professor of Mechanics.

[1] G. G. Stokes, "Dynamic Theory of Diffraction," *Mathematical and Physical Papers* (Cambridge University Press, London, 1883), Vol. II, part 1, pp. 250–290.

[2] Love, *Mathematical Theory of Elasticity* (Dover Publications, New York, 1944), pp. 307–309.

[3] Lord Rayleigh, *Theory of Sound* (Dover Publications, New York, 1945), part II, p. 417.

[4] R. B. Lindsay, J. Acous. Soc. Am. 11, 178–183 (1939).

[5] J. B. Smyth and R. B. Lindsay, J. Acous. Soc. Am. 16, 20–25 (1944).

[6] H. Reissner, "Der senkrechte und schrage Durchtritt einer in einem flüssigen Medium erzeugten ebenen Dilatations Welle durch einen in diesem Medium befindliche planparallele feste Platte," Helv. Phys. Acta. 11, 140–155 (1938).

ρ = mass density
p = frequency in radians/sec.
$h = p/c_\Delta$
$k = p/c_\omega$
σ = normal stress
τ = shearing stress (subscript indicates direction of normal to plane of shear)
l, m, n = direction cosines of normal to advancing wave with respect to x, y, z axes
d = thickness of plates
$P = hn_\Delta d$
$Q = kn_\omega d$
R = reflection coefficient
T = transmission coefficient
$\varphi', \varphi'', \varphi'''$ = incident, reflected and transmitted waves in fluid surrounding the stratified medium
θ_0 = incident angle.

GENERAL EQUATIONS FOR DISPLACEMENTS IN TERMS OF DILATATION AND ROTATION

The dilatation and rotation are defined by the following equations.

$$\Delta = \frac{\partial \xi}{\partial x} + \frac{\partial \eta}{\partial y} + \frac{\partial \zeta}{\partial z} \tag{1}$$

$$\left.\begin{aligned} \omega_x &= \tfrac{1}{2}\left(\frac{\partial \zeta}{\partial y} - \frac{\partial \eta}{\partial z}\right) \\ \omega_y &= \tfrac{1}{2}\left(\frac{\partial \xi}{\partial z} - \frac{\partial \zeta}{\partial x}\right) \\ \omega_z &= \tfrac{1}{2}\left(\frac{\partial \eta}{\partial x} - \frac{\partial \xi}{\partial y}\right) \end{aligned}\right\}. \tag{2}$$

Differentiating Δ with respect to x, y, and z, we obtain the three equations

$$\left.\begin{aligned} \nabla^2\xi &= \frac{\partial \Delta}{\partial x} - 2\frac{\partial \omega_z}{\partial y} + 2\frac{\partial \omega_y}{\partial z}, \\ \nabla^2\eta &= \frac{\partial \Delta}{\partial y} - 2\frac{\partial \omega_x}{\partial z} + 2\frac{\partial \omega_z}{\partial x}, \\ \nabla^2\zeta &= \frac{\partial \Delta}{\partial z} - 2\frac{\partial \omega_y}{\partial x} + 2\frac{\partial \omega_x}{\partial y}. \end{aligned}\right\} \tag{3}$$

The general solution for the displacements in terms of Δ and ω which also satisfies the wave equation for dilation and rotation of harmonic time variation is

$$\left.\begin{aligned} \xi &= -\frac{1}{h^2}\frac{\partial \Delta}{\partial x} + \frac{2}{k^2}\frac{\partial \omega_z}{\partial y} - \frac{2}{k^2}\frac{\partial \omega_y}{\partial z}, \\ \eta &= -\frac{1}{h^2}\frac{\partial \Delta}{\partial y} + \frac{2}{k^2}\frac{\partial \omega_x}{\partial z} - \frac{2}{k^2}\frac{\partial \omega_z}{\partial x}, \\ \zeta &= -\frac{1}{h^2}\frac{\partial \Delta}{\partial z} + \frac{2}{k^2}\frac{\partial \omega_y}{\partial x} - \frac{2}{k^2}\frac{\partial \omega_x}{\partial y}. \end{aligned}\right\} \tag{4}$$

For instance, from the first of Eq. (4)

$$\nabla^2\xi = -\frac{1}{h^2}\frac{\partial}{\partial x}(\nabla^2\Delta) + \frac{2}{k^2}\frac{\partial}{\partial y}(\nabla^2\omega_z) - \frac{2}{k^2}\frac{\partial}{\partial z}(\nabla^2\omega_y)$$

which compared with the first of Eq. (3) results in the wave equations

$$\begin{aligned} (\nabla^2 + h^2)\Delta &= 0, \\ (\nabla^2 + k^2)\omega_z &= 0, \\ (\nabla^2 + k^2)\omega_y &= 0. \end{aligned}$$

GENERAL EQUATIONS FOR STRESSES IN TERMS OF DILATATION AND ROTATION

From the theory of elasticity we have the equations for stress in terms of displacements as follows:

$$\sigma_x = \lambda\Delta + 2G(\partial\xi/\partial x)$$

and two others,

$$\tau_x = G\left(\frac{\partial \eta}{\partial z} + \frac{\partial \zeta}{\partial y}\right)$$

and two others. Hence by substituting from Eq. (4) we obtain the equations for stress in terms of ∇ and ω

$$\sigma_x = \lambda\Delta + 2G\left(-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial x^2} + \frac{2}{k^2}\frac{\partial^2\omega_z}{\partial x\partial y} - \frac{2}{k^2}\frac{\partial^2\omega_y}{\partial x\partial z}\right)$$

$$\sigma_y = \lambda\Delta + 2G\left(-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial y^2} + \frac{2}{k^2}\frac{\partial^2\omega_x}{\partial y\partial z} - \frac{2}{k^2}\frac{\partial^2\omega_z}{\partial x\partial y}\right)$$

$$\sigma_z = \lambda\Delta + 2G\left(-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial z^2} + \frac{2}{k^2}\frac{\partial^2\omega_y}{\partial x\partial z} - \frac{2}{k^2}\frac{\partial^2\omega_x}{\partial y\partial z}\right)$$

$$\begin{aligned} \tau_x = 2G\Bigg(&-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial y\partial z} - \frac{1}{k^2}\frac{\partial^2\omega_z}{\partial x\partial z} - \frac{1}{k^2}\frac{\partial^2\omega_x}{\partial y^2} \\ &+ \frac{1}{k^2}\frac{\partial^2\omega_x}{\partial z^2} + \frac{1}{k^2}\frac{\partial^2\omega_y}{\partial x\partial y}\Bigg) \end{aligned} \tag{5}$$

$$\begin{aligned} \tau_y = 2G\Bigg(&-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial x\partial z} - \frac{1}{k^2}\frac{\partial^2\omega_x}{\partial x\partial y} - \frac{1}{k^2}\frac{\partial^2\omega_y}{\partial z^2} \\ &+ \frac{1}{k^2}\frac{\partial^2\omega_y}{\partial x^2} + \frac{1}{k^2}\frac{\partial^2\omega_z}{\partial y\partial z}\Bigg) \end{aligned}$$

$$\begin{aligned} \tau_z = 2G\Bigg(&-\frac{1}{h^2}\frac{\partial^2\Delta}{\partial x\partial y} - \frac{1}{k^2}\frac{\partial^2\omega_y}{\partial y\partial z} - \frac{1}{k^2}\frac{\partial^2\omega_z}{\partial x^2} \\ &+ \frac{1}{k^2}\frac{\partial^2\omega_z}{\partial y^2} + \frac{1}{k^2}\frac{\partial^2\omega_x}{\partial x\partial z}\Bigg). \end{aligned}$$

TRANSMISSION OF PLANE WAVES THROUGH STRATIFIED SOLID PLATES

Choosing the coordinate axes as shown in Fig. 1, ω_x, ω_y and m will be zero. Considering the nth plate, the dilitation and rotation satisfying the wave equation are

$$\Delta=[\Delta' \exp(ih(l_\Delta x+n_\Delta z))+\Delta'' \exp(ih(l_\Delta x-n_\Delta z))]e^{ipt}$$
$$\omega=[\omega' \exp(ik(l_\omega x+n_\omega z))+\omega'' \exp(ik(l_\omega x-n_\omega z))]e^{ipt} \quad (6)$$

where the primed and double primed quantities represent the advancing and reflected waves. Substituting into Eq. (4) and (5) and omitting the trivial time factor, we obtain the following equations at $x=0$, $z=d$, and $x=0$, $z=0$, expressed in matrix form.

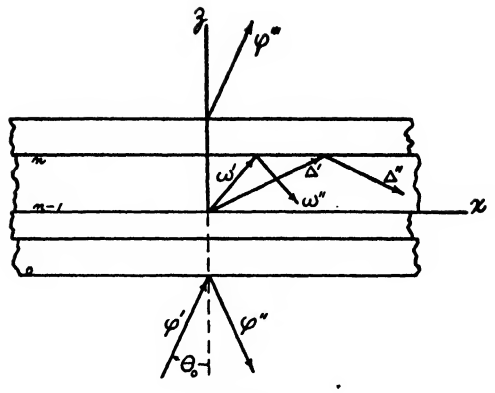


FIG. 1.

$$\begin{vmatrix} \xi_n \\ \zeta_n \\ \sigma_n \\ \frac{\tau_n}{2G} \end{vmatrix} = \begin{vmatrix} \frac{pl_\Delta}{h}\cos P & \frac{ipl_\Delta}{h}\sin P & \frac{2pn_\omega}{k}\cos Q & \frac{i2pn_\omega}{k}\sin Q \\ \frac{ipn_\Delta}{h}\sin P & \frac{pn_\Delta}{h}\cos P & -\frac{i2pl_\omega}{k}\sin Q & -\frac{2pl_\omega}{k}\cos Q \\ (\lambda+2Gn_\Delta^2)\cos P & i(\lambda+2Gn_\Delta^2)\sin P & -4Gl_\omega n_\omega \cos Q & -i4Gl_\omega n_\omega \sin Q \\ il_\Delta n_\Delta \sin P & l_\Delta n_\Delta \cos P & -i(l_\omega^2-n_\omega^2)\sin Q & -(l_\omega^2-n_\omega^2)\cos Q \end{vmatrix} \begin{vmatrix} \Delta'+\Delta'' \\ \Delta'-\Delta'' \\ \omega'-\omega'' \\ \omega'+\omega'' \end{vmatrix}, \quad (7)$$

$$\begin{vmatrix} \xi_{n-1} \\ \zeta_{n-1} \\ \sigma_{n-1} \\ \frac{\tau_{n-1}}{2G} \end{vmatrix} = \begin{vmatrix} \frac{pl_\Delta}{h} & 0 & \frac{2pn_\omega}{k} & 0 \\ 0 & \frac{pn_\Delta}{h} & 0 & -\frac{2pl_\omega}{k} \\ (\lambda+2Gn_\Delta^2) & 0 & -4Gl_\omega n_\omega & 0 \\ 0 & l_\Delta n_\Delta & 0 & -(l_\omega^2-n_\omega^2) \end{vmatrix} \begin{vmatrix} \Delta'+\Delta'' \\ \Delta'-\Delta'' \\ \omega'-\omega'' \\ \omega'+\omega'' \end{vmatrix}. \quad (8)$$

The inversion of Eq. (8) is

$$\begin{vmatrix} \Delta'+\Delta'' \\ \Delta'-\Delta'' \\ \omega'-\omega'' \\ \omega'+\omega'' \end{vmatrix} = \begin{vmatrix} -\frac{4Gl_\omega n_\omega}{D_1} & 0 & -\frac{2pn_\omega}{kD_1} & 0 \\ 0 & -\frac{(l_\omega^2-n_\omega^2)}{D_2} & 0 & \frac{2pl_\omega}{kD_2} \\ -\frac{(\lambda+2Gn_\Delta^2)}{D_1} & 0 & \frac{pl_\Delta}{hD_1} & 0 \\ 0 & -\frac{l_\Delta n_\Delta}{D_2} & 0 & \frac{pn_\Delta}{hD_2} \end{vmatrix} \begin{vmatrix} \xi_{n-1} \\ \zeta_{n-1} \\ \sigma_{n-1} \\ \frac{\tau_{n-1}}{2G} \end{vmatrix} \quad (9)$$

where

$$D_1=-\left[\frac{2pn_\omega}{k}(\lambda+2Gn_\Delta^2)+\frac{pl_\Delta}{h}4Gl_\omega n_\omega\right],$$
$$D_2=-\left[\frac{pn_\Delta}{h}(l_\omega^2-n_\omega^2)-\frac{2pl_\omega}{k}l_\Delta n_\Delta\right]. \tag{10}$$

Thus by substituting Eq. (9) in (7) we obtain the recurrence formula

$$\begin{bmatrix}\xi_n\\ \dot{\zeta}_n\\ \sigma_n\\ \dfrac{\tau_n}{2G}\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&a_{13}&a_{14}\\ a_{21}&a_{22}&a_{23}&a_{24}\\ a_{31}&a_{32}&a_{33}&a_{34}\\ a_{41}&a_{42}&a_{43}&a_{44}\end{bmatrix}\begin{bmatrix}\xi_{n-1}\\ \dot{\zeta}_{n-1}\\ \sigma_{n-1}\\ \dfrac{\tau_{n-1}}{2G}\end{bmatrix} \tag{11}$$

where the expressions for the matrix elements are given in the Appendix.

The continuity of particle velocities and stresses at the interface enables one to express the quantities at the nth plate in terms of corresponding quantities at the first plate. Starting with Eq. (11) we can write

$$\begin{bmatrix}\xi_n\\ \dot{\zeta}_n\\ \sigma_n\\ \dfrac{\tau_n}{2G}\end{bmatrix}=[a_n]\,[a_{n-1}]\cdots[a_1]\begin{bmatrix}\xi_0\\ \dot{\zeta}_0\\ \sigma_0\\ \dfrac{\tau_0}{2G}\end{bmatrix}$$
$$=\begin{bmatrix}A_{11}&A_{12}&A_{13}&A_{14}\\ A_{21}&A_{22}&A_{23}&A_{24}\\ A_{31}&A_{32}&A_{33}&A_{34}\\ A_{41}&A_{42}&A_{43}&A_{44}\end{bmatrix}\begin{bmatrix}\xi_0\\ \dot{\zeta}_0\\ \sigma_0\\ \dfrac{\tau_0}{2G}\end{bmatrix} \tag{12}$$

where $[a_k]$ represents the square matrix of Eq. (11) for the kth plate.

The direction cosines l_Δ, n_Δ, l_ω and n_ω for each plate can be obtained from Snell's law.

$$\frac{\sin\theta_0}{c_0}=\frac{\sin\theta_\Delta}{c_\Delta}=\frac{\sin\theta_\omega}{c_\omega}. \tag{13}$$

Letting θ_1 and θ_2 be the critical angles for the dilation and shear waves respectively, it is apparent that for any plate there will be three regions for θ_0.

(1) When $0\leq\theta_0\leq\theta_1$, both θ_Δ and θ_ω will be real.
(2) When $\theta_1<\theta_0\leq\theta_2$, θ_Δ will be complex and θ_ω real.
(3) When $\theta_2<\theta_0\leq\pi/2$, both θ_Δ and θ_ω will be complex.

As an example, in region (2) we let $\theta_\Delta=\alpha+i\beta$ and obtain $\alpha=\pi/2$ and $\cosh\beta=(c_\Delta/c_0)\sin\theta_0$. Thus we replace $\sin\theta_\Delta$ and $\cos\theta_\Delta$ by $\cosh\beta$ and $i\sinh\beta$ respectively.

If the planes 0 and n terminate in a liquid medium, the shearing stresses τ_0 and τ_n must vanish, resulting in the boundary equation,

$$A_{41}\xi_0+A_{42}\dot{\zeta}_0+A_{43}\sigma_0=0. \tag{14}$$

Substituting Eq. (14) into Eq. (12) with $\tau_0=\tau_n=0$

$$\begin{bmatrix}\dot{\zeta}_n\\ \sigma_n\end{bmatrix}=\begin{bmatrix}\left(A_{22}-\dfrac{A_{21}A_{42}}{A_{41}}\right)&\left(A_{23}-\dfrac{A_{21}A_{43}}{A_{41}}\right)\\ \left(A_{32}-\dfrac{A_{31}A_{42}}{A_{41}}\right)&\left(A_{33}-\dfrac{A_{31}A_{43}}{A_{41}}\right)\end{bmatrix}\begin{bmatrix}\dot{\zeta}_0\\ \sigma_0\end{bmatrix}. \tag{15}$$

To determine the transmission and reflection, we have from Eq. (6) with $\omega=0$ for the liquid medium, the following.

$$\begin{aligned}\dot{\zeta}_n&=c_0n_0\varphi''' & \dot{\zeta}_0&=c_0n_0(\varphi'-\varphi'')\\ \sigma_n&=-c_0^2\rho_0\varphi''' & \sigma_0&=-c_0^2\rho_0(\varphi'+\varphi'')\end{aligned} \tag{16}$$

where φ', φ'' and φ''' represent the incident, reflected, and transmitted waves respectively. Substituting into Eq. (15) the reflection and transmission coefficients become,

$$R=\left|\frac{\varphi''}{\varphi'}\right|^2=\left|\frac{\begin{array}{l}[A_{22}-(A_{21}A_{42}/A_{41})]-[A_{33}-(A_{31}A_{43}/A_{41})]\\ \quad-[A_{23}-(A_{21}A_{43}/A_{41})](c_0\rho_0/n_0)+[A_{32}-(A_{31}A_{42}/A_{41})](n_0/c_0\rho_0)\end{array}}{\begin{array}{l}[A_{22}-(A_{21}A_{42}/A_{41})]+[A_{33}-(A_{31}A_{43}/A_{41})]\\ \quad+[A_{23}-(A_{21}A_{43}/A_{41})](c_0\rho_0/n_0)+[A_{32}-(A_{31}A_{42}/A_{41})](n_0/c_0\rho_0)\end{array}}\right|^2, \tag{17}$$

$$T=1-R. \tag{18}$$

SPECIAL CASES

The following special cases are of interest.

(1) Stratified medium is fluid, thus all shearing stresses are zero. Letting $G=Q=c_\omega=0$, we find that the 1st and 4th columns of $[a_k]$ are zero. This requires that all terms in $[A]$ vanish except the four central terms and we obtain the equations

$$\begin{bmatrix}\dot{\zeta}_n\\ \sigma_n\end{bmatrix}=\begin{bmatrix}A_{22}&A_{23}\\ A_{32}&A_{33}\end{bmatrix}\begin{bmatrix}\dot{\zeta}_0\\ \sigma_0\end{bmatrix} \tag{19}$$

$$R=\left|\frac{A_{22}-A_{33}-A_{23}(c_0\rho_0/n_0)+A_{32}(n_0/c_0\rho_0)}{A_{22}+A_{33}+A_{23}(c_0\rho_0/n_0)+A_{32}(n_0/c_0\rho_0)}\right|^2, \quad (20)$$

$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} \cosh nd & \frac{in}{c\rho}\sinh nd \\ \frac{ic\rho}{n}\sinh nd & \cosh nd \end{vmatrix}. \quad (21)$$

Thus for a single liquid layer we obtain the well-known relation (see Rayleigh, Part II, Eq. (10), p. 88)

$$R=\frac{[(n_1/n_0)(c_0\rho_0/c_1\rho_1)-(n_0/n_1)(c_1\rho_1/c_0\rho_0)]^2}{4\cot^2 h_1n_1d_1+[(n_1/n_0)(c_0\rho_0/c_1\rho_1)+(n_0/n_1)(c_1\rho_1/c_0\rho_0)]^2}. \quad (22)$$

(2) Stratified medium is solid, and the incident ray is normal to the plates. In this case the shears vanish and Eqs. (20), (21), and (22) become applicable with $n_0=n_1=1$.

(3) Transmission through a single solid plate at the critical angles.

(*a*) At the critical angle θ_1 for the dilation wave $l_\Delta=1$ and $n_\Delta=0$. Re-examination of Eqs. (7) and (8) shows that every term in the second row is zero which requires that the terms in the second row of Eq. (11) be also zero. Furthermore, Eq. (15) becomes

$$\begin{vmatrix} \dot{\zeta}_1 \\ \sigma_1 \end{vmatrix} = \begin{vmatrix} 0 & \left(a_{23}-a_{21}\frac{a_{43}}{a_{41}}\right) \\ 0 & \left(a_{33}-a_{31}\frac{a_{43}}{a_{41}}\right) \end{vmatrix} \begin{vmatrix} \dot{\zeta}_0 \\ \sigma_0 \end{vmatrix} \quad (23)$$

and the ratio φ''/φ' reduces to

$$\frac{\varphi''}{\varphi'}=\frac{[a_{23}-a_{21}(a_{43}/a_{41})]-[a_{33}-a_{31}(a_{43}/a_{41})]}{[a_{23}-a_{21}(a_{43}/a_{41})]+[a_{33}-a_{31}(a_{43}/a_{41})]}. \quad (24)$$

Re-evaluating a_{ij}, it can be shown that $[a_{23}-a_{21}(a_{43}/a_{41})]=0$ and $[a_{33}-a_{31}(a_{43}/a_{41})]\neq 0$, thus $\varphi''/\varphi'=-1$ and the transmission is zero.[7]

(*b*) At the critical angle θ_2 for the shear wave $l_\omega=1$ and $n_\omega=0$. We find for this case that the first and third rows in Eq. (11) are zeros. Consequently

$$\frac{\varphi''}{\varphi'}=\frac{(a_{22}/a_{32})-1}{(a_{22}/a_{32})+1} \quad (25)$$

which is equal to -1 only if $a_{22}=0$ and $a_{23}\neq 0$. Thus we find no transmission at θ_2 if the relationship

$$\cosh n_\Delta d=2l_\Delta(c_\omega/c_\Delta) \quad (26)$$

is satisfied, where n_Δ and l_Δ are the direction cosines corresponding to θ_2.

[7] W. C. Schneider and C. J. Burton, J. App. Phys. **20**, 48–58 (1949).

APPENDIX

Matrix elements for Eq. (11) are given with the following substitution.

$$\frac{p}{h}=c_\Delta \qquad n_\Delta=\cos\theta_\Delta$$

$$\frac{p}{k}=c_\omega \qquad n_\omega=\cos\theta_\omega$$

$$G=\rho c_\omega^2 \qquad l_\Delta=\sin\theta_\Delta$$

$$(\lambda+2Gn_\Delta^2)=\rho c_\Delta^2\cos 2\theta_\omega \qquad l_\omega=\sin\theta_\omega$$

$$D_1=-2\rho c_\Delta c_\omega[c_\omega\sin\theta_\Delta\sin 2\theta_\omega+c_\Delta\cos\theta_\omega\cos 2\theta_\omega]$$

$$D_2=[c_\Delta\cos\theta_\Delta\cos 2\theta_\omega+c_\omega\sin\theta_\omega\sin 2\theta_\Delta]$$

$$a_{11}=-\frac{2\rho c_\Delta c_\omega}{D_1}[c_\omega\sin\theta_\Delta\sin 2\theta_\omega\cos P+c_\Delta\cos\theta_\omega\cos 2\theta_\omega\cos Q]$$

$$a_{13}=-\frac{2c_\Delta c_\omega\sin\theta_\Delta\cos\theta_\omega}{D_1}[\cos P-\cos Q]$$

$$a_{21}=-\frac{2ic_\Delta c_\omega\rho}{D_1}[c_\omega\cos\theta_\Delta\sin 2\theta_\omega\sin P-c_\Delta\sin\theta_\omega\cos 2\theta_\omega\sin Q]$$

$$a_{23}=-\frac{2ic_\Delta c_\omega}{D_1}[\cos\theta_\Delta\cos\theta_\omega\sin P+\sin\theta_\Delta\sin\theta_\omega\sin Q]$$

$$a_{31}=-\frac{2c_\Delta^2c_\omega^2\rho^2}{D_1}\sin 2\theta_\omega\cos 2\theta_\omega[\cos P-\cos Q]$$

$$a_{33}=-\frac{2c_\Delta c_\omega\rho}{D_1}[c_\Delta\cos\theta_\omega\cos 2\theta_\omega\cos P+c_\omega\sin\theta_\Delta\sin 2\theta_\omega\cos Q]$$

$$a_{41}=\frac{-i\rho}{D_1}[c_\omega^2\sin 2\theta_\Delta\sin 2\theta_\omega\sin P+c_\Delta^2\cos^2 2\theta_\omega\sin Q]$$

$$a_{43}=-\frac{i}{D_1}[c_\omega\cos\theta_\omega\sin 2\theta_\Delta\sin P-c_\Delta\sin\theta_\Delta\cos 2\theta_\omega\sin Q]$$

$$a_{12}=\frac{-i}{D_2}[c_\Delta\sin\theta_\Delta\cos 2\theta_\omega\sin P+c_\omega\cos\theta_\omega\sin 2\theta_\Delta\sin Q]$$

$$a_{14}=-\frac{2ic_\Delta c_\omega}{D_2}[\sin\theta_\Delta\sin\theta_\omega\sin P+\cos\theta_\Delta\cos\theta_\omega\sin Q]$$

$$a_{22}=\frac{1}{D_2}[c_\Delta\cos\theta_\Delta\cos 2\theta_\omega\cos P+c_\omega\sin\theta_\omega\sin 2\theta_\Delta\cos Q]$$

$$a_{24}=\frac{2c_\Delta c_\omega\sin\theta_\omega\cos\theta_\Delta}{D_2}[\cos P-\cos Q]$$

$$a_{32}=\frac{i\rho}{D_2}[c_\Delta^2\cos^2 2\theta_\omega\sin P+c_\omega^2\sin 2\theta_\Delta\sin 2\theta_\omega\sin Q]$$

$$a_{34}=\frac{2i\rho c_\Delta c_\omega}{D_2}[c_\Delta\sin\theta_\omega\cos 2\theta_\omega\sin P-c_\omega\cos\theta_\Delta\sin 2\theta_\omega\sin Q]$$

$$a_{42}=\frac{\sin 2\theta_\Delta\sin 2\theta_\omega}{2D_2}[\cos P-\cos Q]$$

$$a_{44}=\frac{1}{D_2}[c_\omega\sin\theta_\omega\sin 2\theta_\Delta\cos P+c_\Delta\cos\theta_\Delta\cos 2\theta_\omega\cos Q].$$

The Theory of N Coupled Parallel Antennas*

RONOLD KING
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received July 21, 1949)

The integral-equation theory of coupled antennas developed by King and Harrison, Tai and Bouwkamp for two- and three-coupled antennas is generalized to any number of units symmetrically arranged around a circle. The case of four antennas at the corners of a square is discussed in detail. Application to cage and corner-reflector antennas is indicated.

The analysis of N parallel antennas arranged in line is carried out to determine the driving voltages required to maintain specified currents. It is shown that the inverse problem cannot be carried out in general except as a problem in simultaneous integral equations, but that the special conditions obtaining for half-wave dipoles permit an approximate analysis for the currents with given voltages. This is carried through.

1. GENERAL THEORY

THE conventional method of calculating the mutual impedance of two antennas[1-4] is based on assumed sinusoidally distributed currents on the antennas. A more rigorous method due to King and Harrison[5] and, independently, Bouwkamp,[6] solves the simultaneous integral equations for the currents in two coupled antennas and defines the self- and mutual impedances of the antennas as the coefficients of the currents in the simultaneous voltage equations. This method was improved by C. T. Tai[7] and generalized from two antennas to three arranged at the vertices of an equilateral triangle. Tai extended the analysis to N antennas symmetrically arranged in a circle only in certain special cases. This paper removes the limitations unnecessarily imposed by Tai.

The arrays of antennas to be discussed consist of N geometrically identical units each of half-length h and radius a that satisfy the inequalities, $\beta_0 a = 2\pi a/\lambda_0 \ll 1$, $a \ll h$. The several units may be center-driven by arbitary voltages or center-loaded by different impedances provided the assumed geometrical identity of the units is not significantly disturbed by the driving or load connections or the structure of generators or impedors. In practice, this means either that the driving and load connections are all alike—as when similar open-wire lines are used between all antennas and the generators or loads—or that generators and loads or the cross-sectional dimensions of transmission lines are so small in physical size that generators and loads are equivalent to dimensionless, lumped elements.

The discussion is limited to arrays of *parallel* antennas so that all currents that contribute significantly to the far-zone field are co-directional and, hence, maintain a vector potential that has no component perpendicular to the antennas that are assumed to be parallel to the z-axis of a coordinate system. That is, $\mathbf{A} = \mathfrak{z}\mathbf{A}_z$.

Each antenna is required to have geometrical symmetry with respect to its center, and it must be center-driven or center-loaded. Furthermore, all elements have their centers in the xy-plane, i.e., they are not staggered. These conditions are imposed in order that the currents and vector potentials be even functions of z, the charges and scalar potentials odd functions of z. Thus, with complex quantities set in bold face,

$$\mathbf{I}_z(-z) = \mathbf{I}_z(z); \quad \mathbf{q}(-z) = -\mathbf{q}(z), \tag{1}$$

$$\mathbf{A}_z(-z) = \mathbf{A}_z(z); \quad \boldsymbol{\phi}(-z) = -\boldsymbol{\phi}(z). \tag{2}$$

Let it be assumed at the outset that all generators are "slice" generators, i.e., discontinuities in scalar potential, and all load impedances or tuning reactances lumped or geometrically dimensionless. These assumptions eliminate complications arising from transmission-line end-effects and coupling effects that are necessarily characteristic of particular driving and coupling circuits. Since such junction-effects are localized, they do not alter the formulation of the general problems of coupled antennas, and account may be taken of them later using appropriately designed corrective networks of lumped elements.[8,9] Accordingly, let the driving

* The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (ONR), the Signal Corps of the U. S. Army, and the U. S. Air Force, under ONR Contract N5-ori-76, T. O. 1.

[1] P. S. Carter, "Circuit relations in radiating systems and applications to antenna problems," Proc. I.R.E. 20, 1004 (1932).

[2] G. H. Brown, "Directional antennas," Proc. I.R.E. 25, 78 (1937).

[3] B. Starnecki and E. Fitch, "Mutual impedance of two center-driven parallel aerials," S.R.D.E. Report No. 1023.

[4] K. J. Affanasiev, "Simplifications in the consideration of mutual effects between half-wave dipoles in collinear and parallel orientations," Proc. I.R.E. 34, 635-638 (1946).

[5] R. King and C. W. Harrison, Jr., "Mutual and self-impedance for coupled antennas," J. App. Phys. 15, 481-495 (1944).

[6] C. J. Bouwkamp, "On the theory of coupled antennae," Phillips Res. Rep. 3, 213-226 (1948).

[7] C. T. Tai, "The theory of coupled antennas and its applications," Cruft Laboratory Technical Report No. 12 (March, 1948); "Coupled antennas," Proc. I.R.E. 36, 487-500 (1948).

$$C_a(h,z) \equiv \int_0^h \cos\beta_0 z' \left[\frac{\exp(-j\beta_0 R_1)}{R_1} + \frac{\exp(-j\beta_0 R_2)}{R_2}\right] dz'$$

$$S_a(h,z) \equiv \int_0^h \sin\beta_0 z' \left[\frac{\exp(-j\beta_0 R_1)}{R_1} + \frac{\exp(-j\beta_0 R_2)}{R_2}\right] dz'$$

$$R_1 \equiv ((z-z')^2+a^2)^{\frac{1}{2}}; \quad R_2 \equiv ((z+z')^2+a^2)^{\frac{1}{2}}.$$

[8] R. King, "Theory of antennas driven from two-wire line," Cruft Laboratory Technical Report No. 41 (June 1948). "Antennas and two-wire lines, Part I, J. App. Phys. 20, 832 (1949).

[9] R. King and K. Tomiyasu, "Terminal impedance and generalized two-wire line theory," Cruft Laboratory Technical Report No. 74 (April 15, 1949); Proc. I.R.E. 37, 1134 (1949).

voltage at the center of a typical antenna k be defined as follows:

$$\mathbf{V}_{k0}=\lim_{\delta\to 0}[\phi_k(\delta)-\phi_k(-\delta)]. \quad (3)$$

If the several conditions which have been imposed thus far are satisfied, the vector potential, $\mathbf{A}_z(z)$, on the surface of each of the N antennas satisfies the same one-dimensional wave equation as when the antenna is isolated. For example, on the surface of the kth antenna the equation is:

$$\partial^2\mathbf{A}_{kz}(z)/\partial z^2+\beta_0^2\mathbf{A}_{kz}(z)=j(\beta_0^2/\omega)\mathbf{z}^i\mathbf{I}_{kz}(z)\doteq 0. \quad (4)$$

For simplicity, the term which has the internal impedance per unit length, $\mathbf{z}^i$, as a factor is set equal to zero. For good conductors its effect is negligibly small. If desired, the appropriate terms are easily added to the final solution. The general solution of the homogeneous equation for the range $0\leqq z\leqq h$ is[10]

$$\mathbf{A}_{kz}(z)=-(j/v_0)[\mathbf{C}_k\cos\beta_0 z+\tfrac{1}{2}\mathbf{V}_{k0}\sin\beta_0 z]. \quad (5)$$

The arbitrary constant $\mathbf{C}_k$ must be determined from the boundary condition at $z=h$.

The integral for the vector potential at a point z on the surface of antenna k is as follows

$$\mathbf{A}_{kz}(z)=\frac{1}{4\pi v_0}\int_{-h}^{h}\sum_{i=1}^{N}\mathbf{I}_i(z_i')\frac{\exp(-j\beta_0 R_{ki})}{R_{ki}}dz_i', \quad (6a)$$

where

$$R_{ki}=((z_k-z_i')^2+b_{ki}^2)^{\frac{1}{2}};\quad R_{kk}=((z_k-z_k')^2+a^2)^{\frac{1}{2}}. \quad (6b)$$

Subscripts are used on the coordinate z where necessary to indicate on which antenna the distance from the xy-plane is measured. Note that the primed coordinates locate an element of integration on the axis of an antenna, the unprimed coordinates an element on the surface; b_{ki} is the distance between the axes of antenna i and antenna k.

The substitution of (6a) in (5) gives one of N simultaneous integral equations for the N currents in the N coupled elements of the array. As usual $\zeta_0=v_0/\nu_0\doteq 376.7$ ohms.

$$\int_{-h}^{h}\sum_{i=1}^{N}\mathbf{I}_{iz}(z_i')\frac{\exp(-j\beta_0 R_{ki})}{R_{ki}}dz_i'=\frac{-j4\pi}{\zeta_0}[\mathbf{C}_k\cos\beta_0 z+\tfrac{1}{2}\mathbf{V}_{k0}\sin\beta_0 z];\quad k=1,2,\cdots N. \quad (7)$$

Since the solution of N simultaneous integral equations of this type has not been accomplished even for $N=2$, the only cases for which a solution is available are those in which all the N equations are alike, so that they reduce to only a single equation. Evidently, such a reduction involves restrictions in addition to those already imposed. Mathematically it implies that the current $\mathbf{I}_{kz}(z')$, which is included in the sum in (7), can be factored out in front of the sign of summation and the remaining sum is a constant that does not vary with the numerical value of k. That is, the desired equation must have the following form:

$$\int_{-h}^{h}\mathbf{I}_{kz}(z_k')\mathbf{K}_\Sigma(z,z')dz_k'=\frac{-j4\pi}{\zeta_0}[\mathbf{C}_k\cos\beta_0 z+\tfrac{1}{2}\mathbf{V}_{k0}\sin\beta_0 z], \quad (8a)$$

where

$$\mathbf{K}_\Sigma(z,z')=\sum_{i=1}^{N}\left[\frac{\mathbf{I}_{iz}(z_i')}{\mathbf{I}_{ik}(z_k')}\exp(-j\beta_0 R_{ik})\right]\frac{1}{R_{ik}}$$

is independent of k, (8b)

and

$$R_{ik}=((z_i-z_k')+b_{ik}^2)^{\frac{1}{2}};\quad R_{kk}=((z_k-z_k')^2+a^2)^{\frac{1}{2}}. \quad (8c)$$

Evidently, this is not true in general. It is necessary that both the ratio of currents and the function of R_{ik} be independent of k.

Physically, the reduction of the N simultaneous integral equations to a single equation necessarily means that the N units are electrically and geometrically indistinguishable. These conditions certainly require the antennas to be located at the corners of an N-sided equilateral polygon. If this is true

$$\sum_{i=1}^{N}\exp(-j\beta_0 R_{ik})/R_{ik}$$

is obviously independent of k. The simplest way to make the current ratio in (8b) independent of k is to require all currents to be equal in magnitude and in phase, a condition achieved by identical driving voltages on all units. However, this condition is severer than required. Another, somewhat more general possibility is to have the current in each unit equal in magnitude but differing in phase from the next unit around the polygon by a fixed angle $m\theta$. That is,

$$\mathbf{I}_{iz}(z')=\mathbf{I}_{i-1,z}(z')p^m;\quad m=0,1,2,\cdots N-1, \quad (9a)$$

where

$$p=e^{j\theta};\quad \theta=2\pi/N. \quad (9b)$$

Or, more generally,

$$\mathbf{I}_{iz}(z')=\mathbf{I}_{kz}(z')\mathbf{p}^{(i-k)m}. \quad (10)$$

With (9a) substituted in (7), this reduces to the desired form (8a) with

$$\mathbf{K}_\Sigma(z,z')=\sum_{i=1}^{N}\mathbf{p}^{(i-k)m}\frac{\exp(-j\beta_0 R_{ik})}{R_{ik}}. \quad (11)$$

[10] R. King and D. Middleton, "The cylindrical antenna; current and impedance," Quart. App. Math. 3, 302-335 (1946).

Note that the variable of integration is changed from z_i' to z_k'. Obviously (8c) and (6b) are expressions for the *same* distances. It is readily shown that (11) is independent of k by adding an arbitrary integer q to k. Since the *order* of summation makes no difference, the same integer q may be added to i to give $i'=i+q$. The limits of summation of i' are the same as for i. Thus,

$$\mathbf{K}_\Sigma(z, z') = \sum_{i'=1}^{N} \mathbf{p}^{(i'-k-q)} \frac{\exp(-j\beta_i R_{i'k+q})}{R_{i'k+q}}. \tag{12}$$

Since $i'-k-q=i-q$, and geometrical symmetry makes

$$R_{ik} = R_{i+q,\,k+q}, \tag{13}$$

it follows that (12) is identically the same as (11). Since $\mathbf{K}_\Sigma(z, z')$ is independent of k, it is convenient to choose $k=1$ and define the kernel $\mathbf{K}_\Sigma(z, z')$ as follows:

$$\mathbf{K}_{\Sigma m}(z, z') = \sum_{i=1}^{N} \mathbf{p}^{(i-l)m} \frac{\exp(-j\beta_0 R_{il})}{R_{il}}. \tag{14}$$

Currents of equal magnitude and with a progressive, constant phase difference around the symmetrically located units are possible only when the driving voltages are equal in magnitude and have the same phase relations as the currents. That is

$$\mathbf{V}_{i0} = \mathbf{V}_{k0}\mathbf{p}^{(i-k)m}. \tag{15}$$

Since the constant $\mathbf{C}_k$ in (8a) has $\mathbf{V}_{k0}$ as a factor when it is evaluated, it follows that multiplication of (8a) by $\mathbf{p}^{(i-k)m}$ on both sides changes it to an equation for antenna i instead of for antenna k.

The integral equation for the current in the upper half of each antenna in the symmetrical array is:

$$\int_{-h}^{h} \mathbf{I}(z')\mathbf{K}_{\Sigma m}(z, z')dz' = \frac{-j4\pi}{\zeta_0}[\mathbf{C}\cos\beta_0 z + \tfrac{1}{2}\mathbf{V}_0 \sin\beta_0 z], \tag{16}$$

with $\mathbf{K}_{\Sigma m}(z, z')$ given in (14). This equation is formally exactly like that for two coupled antennas.[7] However, the kernel is more complicated since it involves N terms instead of two. The solution of (16) evidently is the same as for two antennas with $\mathbf{K}_{\Sigma m}(z, z')$ substituted for $\mathbf{K}_s(z, z')$. By mere changes in notation an expression for the current is obtained with an expansion parameter defined in terms of $\mathbf{K}_{\Sigma m}(z, z')$ instead of $\mathbf{K}_s(z, z')$. The current so obtained applies to any one of the N units, say antenna 1. Currents in the other units are obtained by multiplying this current by $\mathbf{p}^{m(i-1)}$ with $i=2, 3, \cdots N$.

It has been shown that the general problem of N-coupled, parallel, identical, non-staggered antennas reduces to a single integral equation that is like the equation for an isolated antenna provided the antennas are all located at the corners of a regular polygon and provided they are individually excited by generators or transmission lines that maintain voltages across their terminals that are equal in magnitude and vary progressively in phase by a constant angle from unit to unit around the polygon. For an array of N antennas the constant phase angle may have any one of the values $m\theta$ where $\theta=2\pi/N$ and m is an integer less than N. Since $m=0$ is a possible value, it is seen that there are N available and different angles for each of which a single integral equation governs all currents. These N possible values of m are called *phase sequences*, as is customary in the well-known method of symmetrical components. They may be defined as follows:

Zero phase sequence:

$$m=0: \quad \mathbf{V}_{10}=\mathbf{V}_{20}=\mathbf{V}_{30}=\cdots=\mathbf{V}_{N0}=\mathbf{V}^{(0)}; \quad \mathbf{I}_{10}=\mathbf{I}_{20}=\mathbf{I}_{30}=\cdots=\mathbf{I}_{N0}=\mathbf{I}^{(0)}. \tag{17a}$$

$2\pi/N$ phase sequence:

$$m=1: \quad \mathbf{V}_{10}=\mathbf{V}^{(1)};\ \mathbf{V}_{20}=\mathbf{p}\mathbf{V}^{(1)};\ \mathbf{V}_{30}=\mathbf{p}^2\mathbf{V}^{(1)};\ \cdots;\ \mathbf{V}_{N0}=\mathbf{p}^{N-1}\mathbf{V}^{(1)}; \quad \mathbf{I}_{10}=\mathbf{I}^{(1)};\ \mathbf{I}_{20}=\mathbf{p}\mathbf{I}^{(1)};\ \mathbf{I}_{30}=\mathbf{p}^2\mathbf{I}^{(1)};\ \cdots;\ \mathbf{I}_{N0}=\mathbf{p}^{N-1}\mathbf{I}^{(1)}. \tag{17b}$$

$4\pi/N$ phase sequence:

$$m=2: \quad \mathbf{V}_{10}=\mathbf{V}^{(2)};\ \mathbf{V}_{20}=\mathbf{p}^2\mathbf{V}^{(2)};\ \mathbf{V}_{30}=\mathbf{p}^4\mathbf{V}^{(2)};\ \cdots;\ \mathbf{V}_{N0}=\mathbf{p}^{2N-2}\mathbf{V}^{(2)}; \quad \mathbf{I}_{10}=\mathbf{I}^{(2)};\ I_{20}=\mathbf{p}^2\mathbf{I}^{(2)};\ I_{30}=\mathbf{p}^4\mathbf{I}^{(2)};\ \cdots;\ \mathbf{I}_{N0}=\mathbf{p}^{2N-2}\mathbf{I}^{(2)}. \tag{17c}$$

$2m\pi/N$ phase sequence:

$$m=m: \quad \mathbf{V}_{10}=\mathbf{V}^{(m)};\ \mathbf{V}_{20}=\mathbf{p}^m\mathbf{V}^{(m)};\ \mathbf{V}_{30}=\mathbf{p}^{2m}\mathbf{V}^{(m)};\ \cdots;\ \mathbf{V}_{N0}=\mathbf{p}^{mN-m}\mathbf{V}^{(m)}; \quad \mathbf{I}_{10}=\mathbf{I}^{(m)};\ \mathbf{I}_{20}=\mathbf{p}^m\mathbf{I}^{(m)};\ \mathbf{I}_{30}=\mathbf{p}^{2m}I^{(m)};\ \cdots;\ \mathbf{I}_{N0}=\mathbf{p}^{mN-m}\mathbf{I}^{(m)}. \tag{17d}$$

$2(N-1)\pi/N$ phase sequence:

$$m=N-1: \quad \mathbf{V}_{10}=\mathbf{V}^{(N-1)};\ \mathbf{V}_{20}=\mathbf{p}^{N-1}\mathbf{V}^{(N-1)};\ \mathbf{V}_{30}=\mathbf{p}^{2N-2}\mathbf{V}^{(N-1)};\ \cdots;\ \mathbf{V}_{N0}=\mathbf{p}^{N^2-2N+1}\mathbf{V}^{(N-1)}; \quad \mathbf{I}_{10}=\mathbf{I}^{(N-1)};\ \mathbf{I}_{20}=\mathbf{p}^{N-1}\mathbf{I}^{(N-1)};\ \mathbf{I}_{30}=\mathbf{p}^{2N}\mathbf{I}^{(N-1)};\ \cdots;\ \mathbf{I}_{N0}=\mathbf{p}^{N^2-2N+1}\mathbf{I}^{(N-1)}. \tag{17e}$$

Note that superscripts on $\mathbf{p}$ are powers, superscripts on $\mathbf{V}$ and $\mathbf{I}$ are written in parentheses and denote phase-sequence numbers. Note also that

$$\mathbf{p}^N = e^{j2\pi} = 1;\ \mathbf{p}^{N/2} = e^{j\pi} = -1; \quad \mathbf{p}^{N/4} = e^{j\pi/2} = j;\ \mathbf{p}^{3N/4} = e^{j3\pi/2} = -j. \tag{18}$$

The current $\mathbf{I}^{(m)}$ in each antenna can be determined independently for each phase sequence with an arbitrary value of driving voltage $\mathbf{V}^{(m)}$ using (16). If each antenna is imagined to be driven by all voltages in all N

phase sequences simultaneously, with the N voltages in each unit in series, the following driving voltages apply:

$$\mathbf{V}_{10}=\mathbf{V}^{(0)}+\mathbf{V}^{(1)}+\mathbf{V}^{(2)}+\cdots+\mathbf{V}^{(N-1)}=\sum_{i=0}^{N-1}\mathbf{V}^{(i)} \tag{19a}$$

$$\mathbf{V}_{20}=\mathbf{V}^{(0)}+\mathbf{p}\mathbf{V}^{(1)}+\mathbf{p}^2\mathbf{V}^{(2)}+\cdots+\mathbf{p}^{N-1}\mathbf{V}^{(N-1)}=\sum_{i=0}^{N-1}\mathbf{p}^i\mathbf{V}^{(i)} \tag{19b}$$

$$\mathbf{V}_{30}=\mathbf{V}^{(0)}+\mathbf{p}^2\mathbf{V}^{(1)}+\mathbf{p}^4\mathbf{V}^{(2)}+\cdots+\mathbf{p}^{2N-2}\mathbf{V}^{(N-1)}=\sum_{i=0}^{N-1}\mathbf{p}^{2i}\mathbf{V}^{(i)} \tag{19c}$$

$$\vdots$$

$$\mathbf{V}_{N0}=\mathbf{V}^{(0)}+\mathbf{p}^{N-1}\mathbf{V}^{(1)}+\mathbf{p}^{2N-2}\mathbf{V}^{(2)}+\cdots+\mathbf{p}^{N-2N+1}\mathbf{V}^{(N-1)}=\sum_{i=0}^{N-1}\mathbf{p}^{Ni-i}\mathbf{V}^{(i)}. \tag{19d}$$

Similarly the resultant currents in the antennas are:

$$\mathbf{I}_{10}=\sum_{i=0}^{N-1}\mathbf{I}^{(i)} \tag{20a}$$

$$\mathbf{I}_{20}=\sum_{i=0}^{N-1}\mathbf{p}^i\mathbf{I}^{(i)} \tag{20b}$$

$$\vdots$$

$$\mathbf{I}_{N0}=\sum_{i=0}^{N-1}\mathbf{p}^{Ni-i}\mathbf{I}^{(i)}. \tag{20c}$$

Since the N voltages $\mathbf{V}^{(i)}$, $i=0, 1, 2, \cdots N-1$, are arbitrary, it is possible to determine a set of N such voltages which satisfy (19a-d) for all possible values of $\mathbf{V}_{10}, \mathbf{V}_{20}, \cdots \mathbf{V}_{N0}$. That is, the general problem of solving for the N currents in N coupled antennas located at the corners of a regular polygon may be carried out by first solving for the N independent currents due to voltages chosen to satisfy the conditions of the N phase sequences. An appropriate superposition of these N solutions yields the currents and impedances in all units due to an arbitrary set of driving voltages.

As a simple illustration, let the general theory be applied to two coupled antennas. In this case $N=2$ and m is restricted to the values 0 and 1. That is, there are two phase sequences, the zero phase sequence previously called the symmetrical case, and the π phase-sequence previously called the antisymmetrical case.

Equations corresponding to (19) and (20) are

$$\mathbf{V}_{10}=\mathbf{V}^{(0)}+\mathbf{V}^{(1)}=\mathbf{V}^{s}+\mathbf{V}^{a}, \tag{21a}$$

$$\mathbf{V}_{20}=\mathbf{V}^{(0)}+\mathbf{p}\mathbf{V}^{(1)}=\mathbf{V}^{(0)}-\mathbf{V}^{(1)}=\mathbf{V}^{s}-\mathbf{V}^{a}. \tag{21b}$$

The corresponding kernels for the integral equation are obtained from (14) with $m=0$ and 1 to be

$$\mathbf{K}_{\Sigma 0}(z, z')\equiv\mathbf{K}_{s}(z, z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}+\frac{\exp(-j\beta_0 R_{12})}{R_{12}} \tag{22a}$$

$$\mathbf{K}_{\Sigma 1}(z, z')\equiv\mathbf{K}_{a}(z, z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}+\frac{\exp(-j\beta_0 R_{12})}{R_{12}}. \tag{22b}$$

These results are the same as previously obtained.[7] The theory for N antennas in terms of N phase sequences is a simple generalization of the theory developed by Tai for two and three antennas with two and three phase sequences. Its application to four antennas follows.

2. FOUR ANTENNAS AT CORNERS OF A SQUARE

The integral equation for the current in the upper half of each of four identical, parallel, non-staggered antennas placed at the corners of a square of side $b_{12}=b$ and diagonal $b_{13}=b\sqrt{2}$ is the same as (1.16). The kernel in this case is obtained from (1.15) with $n=4$. It is

$$\mathbf{K}_{\Sigma m}(z, z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}+\mathbf{p}^m\frac{\exp(-j\beta_0 R_{12})}{R_{12}}+\mathbf{p}^{2m}\frac{\exp(-j\beta_0 R_{13})}{R_{13}}+\mathbf{p}^{3m}\frac{\exp(-j\beta_0 R_{14})}{R_{14}} \tag{1}$$

where

$$\mathbf{p}=e^{j\theta};\ \theta=\pi/2 \tag{2a}$$

so that

$$\mathbf{p}=j;\ \mathbf{p}^2=-1;\ \mathbf{p}^3=-j;\ \mathbf{p}^4=1. \tag{2b}$$

With $n=4$, it follows that

$$m=0, 1, 2, 3. \tag{4}$$

The four phase sequences are illustrated schematically in Fig. 1. They include the zero-phase sequence ($m=0$) with all antennas in phase; the $\pi/4$ or positive phase sequence ($m=1$) with antenna $k+1$ leading antenna k by $\pi/4$, $k=1, 2, 3, 4$; the π-phase sequence ($m=2$) with currents alternating in direction around the square; and the $3\pi/4$, $-\pi/4$, or negative phase sequence with antenna $k+1$ leading antenna k by $3\pi/4$, $k=1, 2, 3, 4$.

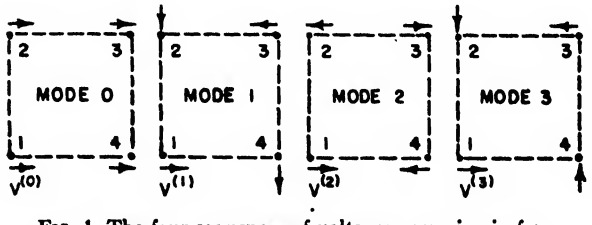


FIG. 1. The four sequences of voltages occurring in four coupled antennas.

The four kernels are

$$m=0:\quad \mathbf{K}_{\Sigma 0}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}+\frac{\exp(-j\beta_0 R_{12})}{R_{12}}+\frac{\exp(-j\beta_0 R_{13})}{R_{13}}+\frac{\exp(-j\beta_0 R_{14})}{R_{14}} \tag{5a}$$

$$m=1:\quad \mathbf{K}_{\Sigma 1}(z,z')=\frac{\exp-(j\beta_0 R_{11})}{R_{11}}+j\frac{\exp(-j\beta_0 R_{12})}{R_{12}}-\frac{\exp(-j\beta_0 R_{13})}{R_{13}}-j\frac{\exp(-j\beta_0 R_{14})}{R_{14}} \tag{5b}$$

$$m=2:\quad \mathbf{K}_{\Sigma 2}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}-\frac{\exp(-j\beta_0 R_{12})}{R_{12}}+\frac{\exp(-j\beta_0 R_{13})}{R_{13}}-\frac{\exp(-j\beta_0 R_{14})}{R_{14}} \tag{5c}$$

$$m=3:\quad \mathbf{K}_{\Sigma 3}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}-\frac{\exp(-j\beta_0 R_{12})}{R_{12}}-\frac{\exp(-j\beta_0 R_{13})}{R_{13}}+j\frac{\exp(-j\beta_0 R_{14})}{R_{14}}. \tag{5d}$$

These expressions can be simplified if use is made of the geometrical condition $R_{12}=R_{14}$ with which (5) becomes

$$m=0:\quad \mathbf{K}_{\Sigma 0}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}+2\frac{\exp(-j\beta_0 R_{12})}{R_{12}}+\frac{\exp(-j\beta_0 R_{13})}{R_{13}} \tag{6a}$$

$$m=1,3:\quad \mathbf{K}_{\Sigma 1}(z,z')=\mathbf{K}_{\Sigma 3}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}-\frac{\exp(-j\beta_0 R_{13})}{R_{13}} \tag{6b}$$

$$m=2:\quad \mathbf{K}_{\Sigma 2}(z,z')=\frac{\exp(-j\beta_0 R_{11})}{R_{11}}-2\frac{\exp(-j\beta_0 R_{12})}{R_{12}}+\frac{\exp(-j\beta_0 R_{13})}{R_{13}}. \tag{6c}$$

If these three kernels are substituted in

$$\Psi_{\Sigma m}(z)\sin\beta_0(h-|z|)=\int_{-h}^{h}\mathbf{K}_{\Sigma m}(z,z')\sin\beta_0(h-z')dz', \tag{7a}$$

the corresponding expansion parameters may be evaluated from

$$\Psi_{\Sigma m}=\left\{\begin{array}{ll}|\Psi_{\Sigma m}(0)|; & \beta_0 h\leqq\pi/2\\ |\Psi_{\Sigma m}(h-\lambda_0/4)|; & \beta_0 h\geqq\pi/2\end{array}\right\} \tag{7b}$$

with

$$\Psi_{\Sigma 0}(z)\sin\beta_0(h-z)=[C_a(h,z)+2C_b(h,z)+C_c(h,z)]\sin\beta_0 h+[S_a(h,z)+2S_b(h,z)+S_c(h,z)]\cos\beta_0 h. \tag{8a}$$

$$\mathbf{I}_{\Sigma 1}(z)\sin\beta_0(h-z)=[C_a(h,z)-C_c(h,z)]\sin\beta_0 h+[S_a(h,z)-S_c(h,z)]\cos\beta_0 h \tag{8b}$$

$$\mathbf{I}_{\Sigma 2}(z)\sin\beta_0(h-z)=[C_a(h,z)-2C_b(h,z)+C_c(h,z)]\sin\beta_0 h+[S_a(h,z)-2S_b(h,z)+S_c(h,z)]\cos\beta_0 h. \tag{8c}$$

The function $C_c(h,z)$ and $S_c(h,z)$ are like the corresponding functions with subscript a defined by Tai[7] with c substituted for a, where $c=b_{13}=b\sqrt{2}$.

With the expansion parameters for the four phase sequences evaluated, the corresponding currents and impedances may be determined just as for two and three antennas.

The general case in which each of the four antennas is driven by an arbitrary voltage may be solved by superposition using the following relations:

$$\mathbf{V}_{10}=\mathbf{V}^{(0)}+\mathbf{V}^{(1)}+\mathbf{V}^{(2)}+\mathbf{V}^{(3)} \tag{9a}$$

$$\mathbf{V}_{20}=\mathbf{V}^{(0)}+j\mathbf{V}^{(1)}-\mathbf{V}^{(2)}-j\mathbf{V}^{(3)} \tag{9b}$$

$$\mathbf{V}_{30}=\mathbf{V}^{(0)}-\mathbf{V}^{(1)}+\mathbf{V}^{(2)}-\mathbf{V}^{(3)} \tag{9c}$$

$$\mathbf{V}_{40}=\mathbf{V}^{(0)}-j\mathbf{V}^{(1)}-\mathbf{V}^{(2)}+j\mathbf{V}^{(3)}. \tag{9d}$$

Alternatively, solving for the phase-sequence voltages,

$$\mathbf{V}^{(0)}=\tfrac{1}{4}(\mathbf{V}_{10}+\mathbf{V}_{20}+\mathbf{V}_{30}+\mathbf{V}_{40}) \tag{10a}$$

$$\mathbf{V}^{(1)}=\tfrac{1}{4}(\mathbf{V}_{10}-j\mathbf{V}_{20}-\mathbf{V}_{30}+j\mathbf{V}_{40}) \tag{10b}$$

$$\mathbf{V}^{(2)}=\tfrac{1}{4}(\mathbf{V}_{10}-\mathbf{V}_{20}+\mathbf{V}_{30}-\mathbf{V}_{40}) \tag{10c}$$

$$\mathbf{V}^{(3)}=\tfrac{1}{4}(\mathbf{V}_{10}+j\mathbf{V}_{20}-\mathbf{V}_{30}-j\mathbf{V}_{40}). \tag{10d}$$

The same equations apply to the currents if I is substituted for V in (9) and (10). By setting

$$\mathbf{V}^{(0)}=\mathbf{I}^{(0)}\mathbf{Z}^{(0)};\ \mathbf{V}^{(1)}=\mathbf{I}^{(1)}\mathbf{Z}^{(1)};\ \mathbf{V}^{(2)}=\mathbf{I}^{(2)}\mathbf{Z}^{(2)};\ \mathbf{V}^{(3)}=\mathbf{I}^{(3)}\mathbf{Z}^{(3)} \tag{11}$$

in (9) and substituting (10) with I written for V in the resulting equations, the following results are obtained:

$$\mathbf{V}_{i0}=\sum_{j=1}^{4}\mathbf{I}_{j0}\mathbf{Z}_{ij};\quad i=1,2,3,4. \tag{12}$$

The evaluation of the self-impedance $\mathbf{Z}_{ii}$ and mutual impedances $\mathbf{Z}_{ij}$ is facilitated if note is taken of the following relations which are a consequence of geometrical and electrical symmetry and the reciprocal theorem:

$$\mathbf{Z}^{(1)}=\mathbf{Z}^{(3)} \tag{13a}$$

$$\mathbf{Z}_{11}=\mathbf{Z}_{22}=\mathbf{Z}_{33}=\mathbf{Z}_{44} \tag{13b}$$

$$\mathbf{Z}_{12}=\mathbf{Z}_{23}=\mathbf{Z}_{34}=\mathbf{Z}_{41}=\mathbf{Z}_{21}=\mathbf{Z}_{32}=\mathbf{Z}_{43}=\mathbf{Z}_{14} \tag{13c}$$

$$\mathbf{Z}_{13}=\mathbf{Z}_{24}=\mathbf{Z}_{31}=\mathbf{Z}_{42}. \tag{13d}$$

With (13a–d) the following impedances are obtained directly:

$$\mathbf{Z}_{11}=\tfrac{1}{4}[\mathbf{Z}^{(0)}+2\mathbf{Z}^{(1)}+\mathbf{Z}^{(2)}] \tag{14a}$$

$$\mathbf{Z}_{12}=\tfrac{1}{4}[\mathbf{Z}^{(0)}-\mathbf{Z}^{(2)}] \tag{14b}$$

$$\mathbf{Z}_{13}=\tfrac{1}{4}[\mathbf{Z}^{(0)}-2\mathbf{Z}^{(1)}+\mathbf{Z}^{(2)}]. \tag{14c}$$

Alternatively, solving for the sequence impedances,

$$\mathbf{Z}^{(0)}=\mathbf{Z}_{11}+2\mathbf{Z}_{12}+\mathbf{Z}_{13} \tag{15a}$$

$$\mathbf{Z}^{(1)}=\mathbf{Z}_{11}-\mathbf{Z}_{13} \tag{15b}$$

$$\mathbf{Z}^{(2)}=\mathbf{Z}_{11}-2\mathbf{Z}_{12}+\mathbf{Z}_{13}. \tag{15c}$$

The symmetry between the phase-sequence impedances and the self- and mutual impedances in (15a–c) and the phase-sequence kernels and the exponential terms in (6a–c) is apparent.

With (14a–c) used in (12) with (13), the currents in the four antennas due to arbitrary driving voltages may be determined.

3. CAGE ANTENNAS

Cage antennas are arrays of N closely spaced parallel antennas excited in the *zero phase sequence* so that *all currents are in phase*. The kernel for an N unit array is

$$\mathbf{K}_{\Sigma 0}(z,z')=\sum_{i=1}^{N}\frac{\exp(-j\beta_0 R_{1i})}{R_{1i}} \tag{1}$$

and the expansion function is

$$\Psi_{\Sigma 0}(z)=\int_{-h}^{h} g(z,z')\mathbf{K}_{\Sigma 0}(z,z')dz', \tag{2}$$

where $g(z,z')=\sin\beta_0(h-z')/\sin\beta_0(h-z)$. This may be expressed as follows:

$$\Psi_{\Sigma 0}(z)=\Psi_{K1}(z)+\Psi_{b12}(z)=\Psi_{b13}(z)+\cdots \tag{3}$$

where

$$\Psi_{K1}(z)=\int_{-h}^{h} g(z,z')\frac{\exp(-j\beta_0 R_{11})}{R_{11}}dz';$$

$$R_{11}=((z-z')^2+a^2)^{\frac{1}{2}} \tag{4a}$$

$$\Psi_{b12}(z)=\int_{-h}^{h} g(z,z')\frac{\exp(-j\beta_0 R_{12})}{R_{12}}dz';$$

$$R_{12}=((z-z')^2+b_{12}^2)^{\frac{1}{2}} \tag{4b}$$

$$\Psi_{b13}(z)=\int_{-h}^{h} g(z,z')\frac{\exp(-j\beta_0 R_{13})}{R_{13}}dz';$$

$$R_{13}=((z-z')^2+b_{13}^2)^{\frac{1}{2}} \tag{4c}$$

$$\Psi_{b1N}(z)=\int_{-h}^{h} g(z,z')\frac{\exp(-j\beta_0 R_{1N})}{R_{1N}}dz';$$

$$R_{1N}=((z-z')^2+b_{1N}^2)^{\frac{1}{2}}. \tag{4d}$$

The distance between centers of antennas 1 and 2 is b_{12}, between antennas 1 and N it is b_{1N}.

For antennas sufficiently closely spaced so that

$$\beta_0 b_{1N}\ll 1, \tag{5}$$

$$\Psi_{b1N}(z_r)=\Psi_{K1}(z_r)-2\ln(b_{1N}/a), \tag{6}$$

where $z_r=0$ for $\beta_0 h\leqq\pi/2$, $z_r=h-\lambda_0/4$ for $\beta_0 h\geqq\pi$. Hence,

$$\Psi_{\Sigma 0}=|\Psi_{\Sigma 0}(z_r)|=|N\Psi_{K1}(z_r)-2\sum_{i=2}^{N}\ln b_{1i}/a| \tag{7a}$$

$$=|N\Psi_{K1}(z_r)-2\ln(b_a b_{13}b_{14}\cdots b_{1N}/a^{N-1})|\cdot \tag{7b}$$

Using this formula for the expansion parameter, the current in and impedance for each of the N antennas forming the cage may be determined just as for a single antenna.

For antennas that are not too long, a satisfactory approximate value of the expansion parameter for a single antenna is

$$\Psi_{K1}=2[\ln 2h/a-1]. \tag{8}$$

Since the contribution to Ψ_{K1} by the imaginary part of $\Psi_{K1}(z_r)$ is not great, an approximate formula for (7b) is obtained by replacing $\Psi_{K1}(z_r)$ by Ψ_{K1} as given by (8). The result is

$$\Psi_{\Sigma 0}\doteq N\Psi_{K1}-2\ln(b_{12}b_{13}b_{14}\cdots b_{1N}/a^{N-1}) \tag{9a}$$

$$\doteq 2[\ln(2h)^N/ab_{12}b_{13}\cdots b_{1N}]. \tag{9b}$$

It is a simple matter to determine the radius a_e of a single antenna carrying N times the current of each unit in the cage and having $1/N$ times the impedance. It is necessary merely to set $a=a_e$ in (8), multiply (8) by N, and equate the result to (9b). Thus solving for a_e

$$a_e=(ab_{12}b_{13}b_{14}\cdots b_{1N})^{1/N}. \tag{10}$$

Therefore, a cage antenna consisting of N *closely spaced*, parallel, identical units arranged at regular intervals around a circle is *approximately* equivalent to a single antenna of much larger radius, in the sense that the total current and its axial distribution are approximately the same, and its impedance is $1/N$ that of the individual units if these are all driven separately.

4. CORNER REFLECTOR ANTENNAS

The obvious application of the general theory of N antennas to the corner reflector antenna using the theorem of images has been carried out by Tai[7] in general, since his analysis did not exclude the case of N antennas when the phases alternate.

5. PARALLEL ARRAYS WITH ALL ELEMENTS DRIVEN; BROADSIDE AND END-FIRE ARRAYS; CIRCUIT PROPERTIES[11]

The most important arrays of identical, parallel, non-staggered, center-driven antennas consist of N units

[11] The three-element broadside array was analyzed by C. W. Harrison, Jr., Proc. I.R.E. 34, 204 (1946) using a method which did not take account of the coupled antennas in the kernel but only in a correction term. By limiting himself to the three-element broadside array, Harrison had to deal with only two simultaneous integral equations which he solved approximately.

spaced at equal intervals b along a straight line as shown in Fig. 2. In one class of arrays, the N currents are made equal in magnitude by appropriate driving voltages, and it is the phase relations which determine the different field characteristics of different arrays. For example, in the broadside array all currents are in phase; in the end-fire array the phase of the currents increases from one antenna to the next by an angle $\beta_0 b$. In another class of arrays, the magnitudes of the currents decrease in a prescribed manner from the center of the array outward in both directions.

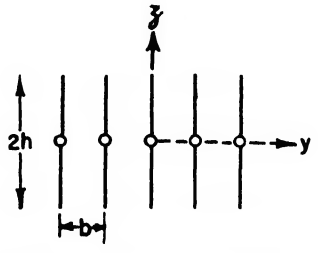


FIG. 2. Five-element parallel array.

Since there are no conditions of symmetry in an array of antennas uniformly spaced along a straight line which make it possible to reduce the N simultaneous integral equations given in (1.7) to a single equation, the conclusion is inevitable that the determination of the N currents and N impedances in terms of arbitrary driving voltages cannot be carried out with available mathematical techniques. Fortunately, an approximate solution of the important inverse problem of determining the N voltages required to maintain N currents with specified relative magnitudes and phases is possible.

The integral equations for the currents in the upper halves ($0 \leqq z \leqq h$) of the N coupled antennas are given by (1.7). They may be expressed as follows:

$$\int_{-h}^{h} \mathbf{I}_{ms}(z_m')\mathbf{K}_p(z_m, z_m')dz_m = \frac{-j4\pi}{\zeta_0}[\mathbf{C}_m \cos\beta_0 z_m + \tfrac{1}{2}\mathbf{V}_{0m} \sin\beta_0 z_m]; \quad m=1, 2, \cdots N, \quad (1a)$$

where the kernel is

$$\mathbf{K}(z_m, z_m') = \sum_{i=1}^{N} \frac{\mathbf{I}_{is}(z')}{\mathbf{I}_{ms}(z_m')} \frac{\exp(-j\beta_0 R_{mi})}{R_{mi}}; \quad m=1, 2, \cdots N, \quad (1b)$$

and where

$$R_{mi} = ((z_m - z_i')^2 + b_{mi}^2)^{\frac{1}{2}}; \; R_{mm} = ((z_m - z_m')^2 + a^2)^{\frac{1}{2}}. \quad (1c)$$

The distance between the centers of antenna m and antenna i is $b_{mi} = |m-i|b$, where b is the equal distance between adjacent units.

The N kernels $\mathbf{K}_p(z_m, z_m')$ are functions of the distributions of current in all antennas and of the distances R_{mi}. They are not independent. However, an approximate solution of (1a) may be obtained by assuming that the *distributions* of current as functions of z' are the same in all N units, but without thereby restricting the magnitudes or the phases of the input currents at $z=0$. It is readily agreed that this is a reasonable assumption if the antennas are identical. Indeed, it has already been shown in the analysis of two coupled antennas[7] driven either in phase or 180 degrees out of phase that the distributions of current do not differ significantly from each other or from the distribution along an isolated antenna provided $\beta_0 h$ does not greatly exceed $\pi/2$. When $\beta_0 h$ is near π the distributions differ considerably for very closely spaced antennas, but as the separation is increased, the currents become more and more nearly alike. Let it be assumed, therefore, that it is a good approximation to set

$$\frac{\mathbf{I}_{is}(z_i')}{\mathbf{I}_{ms}(z_m')} \doteq k_{mi} \exp(j\theta_{mi}), \quad (2)$$

where k_{mi} is the specified real ratio of magnitudes and δ_{mi} the specified phase difference between the currents in antennas k and m except when the antennas are close together *and* have half-lengths h near $\lambda_0/2$.

Substitution of (2) in (1a) gives:

$$\mathbf{K}_p(z_m, z_m') = \sum_{i=1}^{N} k_{mi} \exp(j\theta_{mi}) \frac{\exp(-j\beta_0 R_{mi})}{R_{mi}}; \quad m=1, 2, \cdots N. \quad (3)$$

The N kernels defined in (3) are not alike. They are not and cannot be made independent of m since for geometrical reasons,

$$R_{mi} \neq R_{m+q,\, i+q}, \quad (4)$$

where q is an integer for all values of m and q from 1 to N.

However, since all quantities in (3) are assumed to be known for all values of m, the N kernels can be evaluated independently. In particular, (1a) can be solved separately for $\mathbf{I}_{ms}(z)$ for each value of m. Thus, it is seen that the restrictions on the current contained in (2) have reduced the N *simultaneous* integral equations for determining N independent currents, to N *independent* integral equations for determining the N voltages $\mathbf{V}_{m0}$ that are required to maintain the N specified currents.

Since (1a) differs from the corresponding equation for two antennas only in the number of terms in the kernel, the formal solution for the current and the impedance as a function of the driving voltage may be carried out directly.[7] The expression for the input impedance of antenna m in the presence of $N-1$ parallel antennas all so driven that they have pre-assigned currents is like that for two antennas[7] with additional P

and **Q** terms each multiplied by the appropriate ratio factor and phase factor for the current. Thus,

$$(\mathbf{Z}_m)_{in}=\frac{-j\zeta_0\Psi_{pm}}{2\pi}\left\{\frac{\Psi_{pm}\cos\beta_0 h+\mathbf{F}_1(h)+\mathbf{P}_{1z}(h)}{[\Psi_{pm}+\mathbf{F}_1(0)+\mathbf{P}_{1z}(0)]\sin\beta_0 h-[\mathbf{G}_1(0)+\mathbf{Q}_{1z}(0)]\cos\beta_0 h+[\mathbf{G}_1(h)+\mathbf{Q}_{1z}(h)]}\right\}, \tag{5}$$

where

$$\Psi_{pm}=|\Psi_{pm}(0)|\,;\quad \beta_0 h\leqq\pi/2 \tag{6a}$$

$$\Psi_{pm}=|\Psi_{pm}(h-\lambda_0/4)|\,;\ \beta_0 h\geqq\pi/2 \tag{6b}$$

$$\Psi_{pm}(z)\sin\beta_0(h-z)=[C_a(h,z)+\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\times\exp(j\theta_{mi})C_{bmi}(h,z)]\sin\beta_0 h-[S_a(h,z)+\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_{mi})S_{bmi}(h,z)]\cos\beta_0 h. \tag{6c}$$

Also,

$$\mathbf{P}_1(z)=\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_{mi})\mathbf{P}_{1mi}(z) \tag{7a}$$

$$\mathbf{Q}_1(z)=\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_{mi})\mathbf{Q}_{1mi}(z) \tag{7b}$$

$$\mathbf{P}_{1mi}(z)=-\int_{-h}^{h}\mathbf{F}_{1z'}\frac{\exp(-j\beta_0 R_{mi})}{R_{mi}}dz'=-C_{bmi}(h,z)+E_{bmi}(h,z)\cos\beta_0 h \tag{8a}$$

$$\mathbf{Q}_{1mi}(z)=-\int_{-h}^{h}\mathbf{G}_{1z'}\frac{\exp(-j\beta_0 R_{mi})}{R_{mi}}dz'=-S_{bmi}(h,z)+E_{bmi}(h,z)\sin\beta_0 h \tag{8b}$$

$$R_{mi}=((z_m-z_i')^2+b_{mi}^2)^{\frac{1}{2}}. \tag{8c}$$

$\mathbf{F}_1(0)$, $\mathbf{F}_1(h)$, $\mathbf{G}_1(0)$ and $\mathbf{G}_1(h)$ as well as $C(h,z)$, $S(h,z)$ and $E(h,z)$ are as in reference 7 with Ψ_{pm} substituted for Ψ. Note that b_{mi} is the distance between centers of antennas m and i. Note that when $N=2$, $\theta_{mi}=0$, $k_{mi}=1$, (5) reduces to the previous formula[7] for two symmetrically driven antennas. Alternatively, when $N=2$, $\theta_{mi}=\pi$, $k_{mi}=1$, (5) reduces to the formula for two antisymmetrically driven antennas.

By evaluation of the functions $\mathbf{P}_{1z}$ and $\mathbf{Q}_{1z}$ together with the expansion parameter Ψ_{pm} for *each* value of m, the input impedance of each antenna in the presence of the others may be evaluated. Note that, in general, the antennas are in identical pairs except for the central unit for N odd, so that the number of different values of these functions is $N/2$ for N even, $(N+1)/2$ for N odd. Multiplication of the impedances so determined by the appropriate, assigned, input current gives the driving voltages required to maintain these currents.

In practice, antennas in parallel arrays usually have electrical half-lengths $\beta_0 h$ that are $\pi/2$ or near $\pi/2$. Significantly, this is precisely the value of $\beta_0 h$ for which the initial assumption of identical *distributions* (but not magnitudes or phases) of current is best satisfied and for which a maximum simplification of the intricate formula (5) is possible. With $\beta_0 h=\pi/2$, (5) becomes:

$$(\mathbf{Z}_m)_{in}=\frac{j\zeta_0}{4\pi}\left\{\frac{C_a(h,h)+\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_{mi})C_{bmi}(h,h)}{1-\frac{1}{2\Psi_{pm}}[C_a(h,0)+S_a(h,h)-E_a(h,h)+\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_m)\{C_{bmi}(h,0)+S_{bmi}(h,h-E_{bmi}(h,h)\}]}\right\}. \tag{9}$$

Except for antennas that are extremely closely spaced, the expansion parameter Ψ_{pm} cannot differ greatly from the value Ψ_{K1} for each of the antennas when isolated. In most driven arrays of the types considered in this section, b is not less than $\lambda_0/4$ and frequently is $\lambda_0/2$. Since the term with $1/2\Psi_{pm}$ as a factor in the denominator of (9) necessarily is small compared with unity for antennas of small radius it may be neglected. That this is indeed a good approximation already has been demonstrated for two coupled antennas[7] with $\beta_0 h=\pi/2$. If this term is neglected, (9) reduces to

$$(\mathbf{Z}_m)_{in}=\frac{j\zeta_0}{4\pi}[C_a(h,h)+\sum_{\substack{i=1\\ i\neq m}}^{N} k_{mi}\exp(j\theta_{mi})C_{bmi}(h,h)];\quad m=1,2,\cdots N. \tag{10}$$

Multiplication of (10) by $\mathbf{I}_{m0}$, noting that

$$\mathbf{I}_{i0}=\mathbf{I}_{m0}k_{mi}\exp(j\theta_{mi}), \tag{11}$$

reveals that (10) is equivalent to

$$\begin{aligned}
V_{10}&=I_{10}Z_{11}+I_{20}Z_{12}+I_{30}Z_{13}+\cdots+I_{N0}Z_{1N},\\
V_{20}&=I_{10}Z_{21}+I_{20}Z_{22}+I_{30}Z_{23}+\cdots+I_{N0}Z_{2N},\\
&\cdots\\
V_{N0}&=I_{10}Z_{N1}+I_{20}Z_{N2}+I_{30}Z_{N3}+\cdots+I_{N0}Z_{NN}, \qquad (12)
\end{aligned}$$

where

$$Z_{11}=Z_{22}=Z_{33}=\cdots Z_{NN}=\frac{j\zeta_0}{2\pi}C_a(h,h)=Z_0 \qquad (13)$$

$$Z_{12}=Z_{21}=\frac{j\zeta_0}{2\pi}C_{b12}(h,h) \qquad (14a)$$

$$Z_{13}=Z_{31}=\frac{j\zeta_0}{2\pi}C_{b13}(h,h) \qquad (14b)$$

$$\cdots$$

$$Z_{1N}=Z_{N1}=\frac{j\zeta_0}{2\pi}C_{b1N}(h,h). \qquad (14c)$$

Note that the self-impedances Z_{ii} of the coupled antennas are all the same as the self-impedance Z_0 of an isolated antenna. Moreover, the mutual impedances Z_{ij} are the same as if only antennas i and j were present. Therefore, it may be concluded that for parallel arrays in which antennas of half-length near $h=\lambda_0/4$ are coupled, the self-impedance of an isolated antenna may be used for each antenna and the mutual impedance between any pair taken to be the same as for that pair when isolated. Accordingly, the self-impedance and the mutual impedance may be obtained from first-order curves of Z_{e1},[7] for the appropriate value of $\Omega=2\ln 2h/a$.

It is significant to note that in spite of the initial *general* statement to the contrary, the Eqs. (13) may be solved for either the driving voltages given the current or the currents given the driving voltages. However, this is true only approximately and only in the special case of antennas for which $\beta_0 h=\pi/2$ and Ψ_{12} is not too small.

The simple method of analyzing N coupled antennas in terms of self-impedances of isolated antennas and mutual impedances of two coupled antennas obviously is *not valid* for antennas with half-lengths that exceed $\lambda_0/4$ considerably. Specifically, when $\beta_0 h=\pi$, (5) becomes

$$(Z_m)_{in}=\frac{j\zeta_0\Psi_{pm}^2}{2\pi}\times\left\{\frac{1-(1/\Psi_{pm})[F_1(h)+P_1(h)]}{[G_1(0)+Q_{12}(0)+G_1(h)+Q_{12}(h)]}\right\}. \qquad (15)$$

Since the term involving the most significant contributions from the coupled antennas now appear in the factor Ψ_{pm}^2 and in the terms in Q in the denominator, it is not possible to express $V_{m0}=I_{m0}(Z_m)_{in}$ as a simple sum of self- and mutual terms. $(Z_m)_{in}$ can be evaluated from (16) and V_{m0} evaluated for each antenna if the currents are known. The reverse is not possible.

For the broadside array, $k_{mi}=1$, $\theta_{mi}=0$ with $\beta_0 b=\pi$. For the end-fire array, $\theta_{mi}=\pm(m-i)\beta_0 b$ with $\beta_0 b=\pi/2$ for the unilateral end-fire, π for bilateral end-fire. The analysis in this section has demonstrated that common assumptions, such as that equal driving voltages will produce equal currents in all units of a broadside array, are erroneous except for the two-element array. The different impedances of the differently situated units require different voltages if the currents are to be the same.

Conventional methods of driving parallel arrays in which currents in the units are assumed to be respectively in phase (broadside array) or alternately 180 degrees out-of-phase (bilateral end-fire array) are shown in Fig. 3 and Fig. 4 for three-element arrays. The transmission line connecting the units in the broadside array is assumed to be *spiraled* and *not crossed-over* as shown in the figure so that it is the equivalent of a smooth line one-half wave-length long. Actually, the currents in the three antennas in Figs. 3 and 4 are neither all equal in magnitude nor all in phase (broadside) or alternately 180 degrees out-of-phase (bilateral end-fire) for two reasons. These are, (1) the input impedance of the central unit differs from that of the two outer units since this unit is not in the same position relative to the other antennas in the array; (2) the transmission-line end-effects and coupling effects are different for the central unit from what they are for the two outer units.

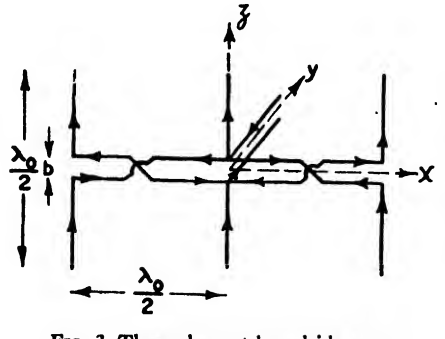


FIG. 3. Three-element broadside array.

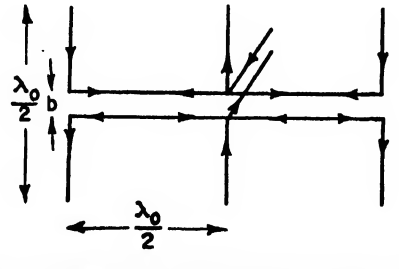


FIG. 4. Three-element bidirectional end-fire array.

Let the problem of transmission-line end-effects and coupling effects between each antenna and the feeding line be sidestepped in this paper by assuming that the distance b between the conductors of the line is sufficiently small to make these effects negligible. If this is done, formulas for the ratio of the current in antenna 1 or 3 (which have equal currents) to the current in the central unit 2 are obtained readily for $\beta_0 h = \pi/2$ since the simple formula (13) with $N=3$ is a good approximation. With negligible terminal-zone effects, the two sections of line of length $\lambda_0/2$ are one-to-one transformers so that the same voltage is applied to each of the three units. That is

$$\mathbf{V}_{30}=\mathbf{V}_{20}=\mathbf{V}_{10}. \tag{16}$$

Moreover, since antennas 1 and 3 are geometrically and electrically identical, it follows that

$$\mathbf{I}_{30}=\mathbf{I}_{10};\ \mathbf{Z}_{12}=\mathbf{Z}_{23}. \tag{17}$$

With (13) and (12) the following relations are obtained directly when $N=3$:

$$\mathbf{V}_{10}=\mathbf{V}_{20}=\mathbf{I}_{10}(\mathbf{Z}_{11}+\mathbf{Z}_{13})+\mathbf{I}_{20}\mathbf{Z}_{12}=2\mathbf{I}_{10}\mathbf{Z}_{12}+\mathbf{I}_{20}\mathbf{Z}_{22}. \tag{18}$$

The solution of (18) for the ratio of currents is

$$\mathbf{I}_{20}/\mathbf{I}_{10}=(\mathbf{Z}_{11}+\mathbf{Z}_{13}-2\mathbf{Z}_{12})/(\mathbf{Z}_{22}-\mathbf{Z}_{12}). \tag{19}$$

The expressions for the impedances are

$$\mathbf{Z}_{3in}=\mathbf{Z}_{1in}\equiv\frac{\mathbf{V}_{10}}{\mathbf{I}_{10}}=\mathbf{Z}_{11}+\mathbf{Z}_{13}+\mathbf{Z}_{12}\left[\frac{\mathbf{Z}_{11}+\mathbf{Z}_{13}-2\mathbf{Z}_{12}}{\mathbf{Z}_{22}-\mathbf{Z}_{12}}\right] \tag{20}$$

$$\mathbf{Z}_{2in}\equiv\frac{\mathbf{V}_{20}}{\mathbf{I}_{20}}=\mathbf{Z}_{22}+2\mathbf{Z}_{12}\left[\frac{\mathbf{Z}_{22}-\mathbf{Z}_{12}}{\mathbf{Z}_{11}+\mathbf{Z}_{12}-2\mathbf{Z}_{12}}\right]. \tag{21}$$

If all antennas are identical and have no series impedances,

$$\mathbf{Z}_{33}=\mathbf{Z}_{22}=\mathbf{Z}_{11}=\mathbf{Z}_{s1}. \tag{22}$$

In this case it is evident that $\mathbf{I}_{30}=\mathbf{I}_{10}$ must differ considerably from $\mathbf{I}_{20}$, and $\mathbf{Z}_{3in}=\mathbf{Z}_{1in}$ from $\mathbf{Z}_{2in}$ since $\mathbf{Z}_{13}$ and $\mathbf{Z}_{12}$ are far from the same. For example,

$$\Omega=10,\ b=\lambda_0/2,\ \mathbf{Z}_{s1}\doteq 73+j31;\quad \mathbf{Z}_{12}=-13-j26;\ \mathbf{Z}_{13}=3+j16. \tag{23}$$

It is possible to make the currents all equal when the driving voltages are the same by connecting an impedance in series with the central unit, antenna 2. Specifically, let

$$\mathbf{Z}_{33}=\mathbf{Z}_{11}=\mathbf{Z}_{s1};\ \mathbf{Z}_{22}=\mathbf{Z}_{s1}+\mathbf{Z}_2. \tag{24}$$

Upon setting the currents equal in (19) and using (24), the result is,

$$\mathbf{Z}_2=\mathbf{Z}_{13}-\mathbf{Z}_{12}. \tag{25}$$

It is seen that $\mathbf{Z}_2$ is quite a large impedance with a significant real part so that appreciable ohmic losses are involved which lower the efficiency of transmission. Since the electromagnetic field of a three-element array in which the current in the central unit is greater than that in the outer units may be more desirable than when the currents are all equal, it is seldom advantageous to use $\mathbf{Z}_2$ as specified in (25) in series with the central unit. An alternative procedure is to connect reactances in series with the *outer* units in order to make the phases of all three units the same. That is, in place of (24), let

$$\mathbf{Z}_{33}=\mathbf{Z}_{11}=\mathbf{Z}_{s1}+jX_1;\ \mathbf{Z}_{22}=\mathbf{Z}_{s1}. \tag{26}$$

The values in (16) now must be substituted in (20) and X_1 adjusted to make the expression real. That is,

$$\frac{\mathbf{I}_{20}}{\mathbf{I}_{10}}=\frac{(R_{s1}+R_{13}-2R_{12})+j(X_1+X_{s1}+X_{13}-2X_{12})}{(R_{s1}-R_{12})+j(X_{s1}-X_{12})} \tag{27}$$

must be real. This is true when

$$\frac{X_1+X_{s1}+X_{13}-2X_{12}}{R_{s1}+R_{13}-2R_{12}}=\frac{X_{s1}-X_{12}}{R_{s1}-R_{12}}, \tag{28a}$$

or when

$$X_1=\left(\frac{X_{s1}-X_{12}}{R_{s1}-R_{12}}\right)\times(R_{s1}+R_{13}-2R_{12})-X_{s1}-X_{13}+2X_{12}. \tag{28b}$$

Evidently

$$\frac{\mathbf{I}_{20}}{\mathbf{I}_{10}}=\frac{R_{s1}+R_{13}-2R_{12}}{R_{s1}-R_{12}}. \tag{29}$$

This ratio may be considerably greater than one.

Similar analyses may be carried out for arrays of any number of parallel units.

A Dimensional Analysis of Metal Cutting*

D. C. Drucker† and H. Ekstein‡
Armour Research Foundation, Chicago, Illinois
(Received June 30, 1949)

Dimensional analysis is used in a study of the behavior of metal when it is cut. Present theories are described and some additional significant variables are found. The analysis provides a basis for the interpretation of known experimental results, and should serve as a guide for future experimental and theoretical investigations.

INTRODUCTION

IF a problem is so complicated that a sufficiently comprehensive theory cannot be developed at the time, a very useful approach to a solution is to write down the independent variables of the problem and form their dimensionless combinations.[1] The significance of each independent dimensionless variable can then be determined from prior knowledge, reasoning, or experiment, and usually many of the dimensionless combinations can be ruled out as insignificant, that is, having no noticeable influence upon the result. Then, if the important independent combinations are the same, all dependent dimensionless variables will also be the same for both experimental and production runs. This means that practical predictions, such as of power requirement and chip continuity or surface finish, can be made from a limited set of experiments or field information without actually testing or ever having tested the proposed set-up. As the dimensionless combinations are not unique, a rearrangement is often worthwhile, for in this way the number of physically important variables may be materially reduced. This procedure has had widespread application in fluid mechanics and associated problems and is well standardized. Its use in mechanical problems is increasing as its value becomes more evident.[2]

Another, and more subtle, use of dimensional analysis is to check the physical completeness of a theory before embarking on the arduous task of grinding out mathematical formulas and numerical calculations. This, too, is illustrated in the following discussion and extension of present theories of metal cutting.

IDEALIZED METAL CUTTING

All theory must involve approximation to, or idealization of, actual behavior. Metal cutting offers a very difficult problem even for the two-dimensional case, Fig. 1, and the description proposed by Piispanen[3] and independently by Ernst and Merchant[4] was a tremendous forward step. The chip is assumed to come off continuously and there is no built-up edge. The fundamental action is a plastic shearing of each plane as it passes OA, the final appearance being similar to that which would be obtained with an inclined deck of very thin cards.

The resultant force R exerted by the tool on the chip per unit width of cut, Fig. 2, its horizontal or power component H, the shear angle φ, the chip thickness, and so forth, are dependent variables determined by the properties of the material being cut and by other independent variables. No one has yet devised a general theory which takes all into account. The major question is which variables are most important and how do they enter. In the simplest theory it is assumed that the independent variables are only the rake angle α, the depth of cut t, the cutting speed V, a constant friction angle between tool and chip β, and a fixed yield stress in shear τ_y. Once this initial step has been taken, R, φ, H, and so forth can be found; Ernst and Merchant have written down the results. However, a dimensional analysis will show that despite its great value there are practical physical limitations for any theory, elaborate or simple, of the form

$$\varphi \text{ or } H \text{ a function of } \alpha, t, V, \beta, \tau_y.$$

φ is dimensionless as are α, and β, but H is a force per

* Based on a research program sponsored by Shell Oil Company, Inc., and performed at Armour Research Foundation, Chicago, Illinois.

† Formerly on the staff of Armour Research Foundation, Chicago, Illinois. Now Associate Professor of Engineering, Brown University, Providence, Rhode Island.

‡ Senior Physicist, Physics Research Department, Armour Research Foundation, Chicago, Illinois.

[1] The reviewer of this paper brought to the authors' attention the excellent article by F. F. P. Bisacre and G. H. Bisacre, "Life of carbide-tipped turning tools," Proc. Inst. Mech. Eng. War Emergency Issue No. 35, 157, 452–469 (1947). A dimensional analysis is included which in certain respects is more complete than ours. However the approach and the objective are quite different and a few variables important here are not included. The reviewer also pointed out that Dr. B. T. Chao has submitted a paper to the Institution of Mechanical Engineers in which, following Bisacre and Bisacre, considerable use is made of $K/Vc\rho t$.

[2] For a particular example in which there are very many independent variables but only one is found to be of real importance in the practical range, see D. C. Drucker and H. Tachau, "A new design criterion for wire rope," J. App. Mech. XII, No. 1, A-33–A-38 (1945).

[3] V. Piispanen, "Lastunmodostumisen teoriaa (Theory of chip formation)," Teknillinen Aikakauslehti XXVII, No. 9, 315–322 (1937).

[4] H. Ernst and M. E. Merchant, "Chip formation, friction, and finish," The Cincinnati Milling Machine Company (1940); (also "Surface treatment of metals," Symposium, Trans. A.S.M.E., 299–378 (1941); M. E. Merchant, "Mechanics of the metal cutting process"; I. "Orthogonal cutting and a type 2 chip," J. App. Phys., XVI, No. 5, 267–275 (1945); II. "Plasticity conditions in orthogonal cutting," J. App. Phys., XVI, No. 6, 318–324 (1945).

unit length, t is a length, V is a length over time, and τ_y is a force divided by length squared. The first step is the formation of the independent dimensionless combinations of the independent variables. It is easily seen that only α and β remain. V alone contains time and τ_y alone contains units of force so that a dimensionless combination cannot be formed with either, and both must be eliminated. Then t remains as the only independent variable involving length, and it must also be dropped. Therefore, any dimensionless quantity containing a dependent variable must turn out to be a function of α and β only, for example,

$$\varphi = \varphi(\alpha, \beta)$$

and

$$H/t\tau_y = H/t\tau_y(\alpha, \beta).$$

This is illustrated by the formulas given by Ernst and Merchant for these conditions, for example,

$$\varphi = 45° - (\beta - \alpha)/2.$$

More elaborate theories are possible and other results can be obtained, but if they are based upon the same five independent quantities only, no formula developed can contain either the speed or the depth of cut as independent variables. As both are known to be important in several ways, such theories, although very useful, are definitely incomplete. Also, the predictions for different materials appear too similar. The calculated shear angle φ will be the same if the friction and rake angles are the same, no matter what the properties of the material.

The most general theory now in use includes the material to the extent of considering the variation of the yield stress in shear with the normal stress σ on the shear plane. The change is generally negligible at ordinary stress magnitudes, but the stresses encountered in metal cutting are extremely high, exceeding 200,000 p.s.i. for mild steel. If the relation is linear, of the form $\tau_y = \tau_0 + k\sigma$, the previous dimensional analysis must be modified by replacing τ_y by the new independent variables τ_0 and k. As k is dimensionless, the result can be written immediately as

$$\phi = \phi(\alpha, \beta k)$$
$$H/t\tau_0 = H/t\tau_0(\alpha, \beta, k)$$

which is shown by Merchant's modified expressions,[4] for example

$$\phi = C/2 - (\beta - \alpha)/2,$$

where C is arc *ctnk*. There is still no influence of depth or speed of cut.

NEW CONCEPTS AND VARIABLES

The necessity for new concepts and the accompanying variables is thus made clear by the dimensional reasoning. It is known that as the speed is decreased or as the depth is increased, the chip will gradually change from the continuous type to the segmental. Present theories therefore cannot indicate under what conditions they break down, that is, when the picture they postulate is no longer valid. Furthermore, the shear angle and the specific cutting force generally do vary with V and t, which was shown to be incompatible with any theory using only the variables mentioned above. Therefore, there is a strong suggestion that a variable T_0 with the dimension of time be introduced. Such an independent variable with physical meaning would include the cutting speed by adding the dimensionless group VT_0/t. T_0 could be a time required for the material to fracture. It would not be valid to introduce such a variable unless T_0 would be of the order of magnitude of the essential time interval in the cutting operation. However, a simple calculation shows that the time in which the shearing action takes place in actual cutting is extremely short and furthermore the time in the ideal case, Fig. 1, is zero. Thus, dimensional analysis can lead to new and useful concepts; in metal cutting the transition from the continuous to the discontinuous chip apparently is governed by the time available for fracture. The transition value of VT_0/t will, of course, depend on the shear strain, which in turn depends upon α and β only. If the additional variables which will be considered later are of minor significance, then the dimensional analysis shows that a given percentage increase in t requires the same percentage increase in V to obtain the same degree of chip continuity and, therefore, approximately the same surface finish.

Another important omission in the idealized picture of cutting is the lack of prediction of any effect of depth of cut on the specific cutting force. It is always observed in practice that when t is small, the value of $H/t\tau_0$ rises very sharply as t decreases. This fact leads to the consideration of lengths which might be significant in cutting. The nose radius of the tool r_N, which was assumed as zero, could account for this phenomenon qualitatively and would appear in the variable r_N/t.

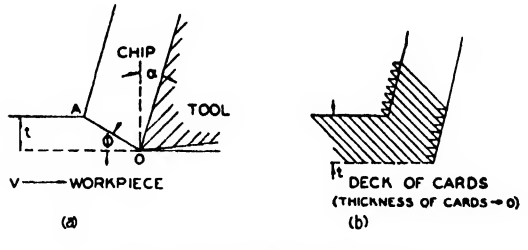


FIG. 1. Two-dimensional cutting.

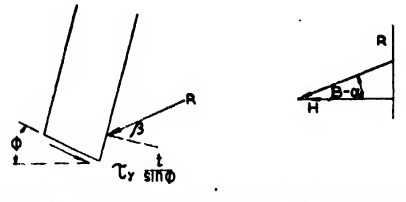


FIG. 2. Forces on chip of unit width (perpendicular to paper).

TABLE I

Variable	Units	Description
c	$FLM^{-1}\theta^{-1}/J$	specific heat, for example, B.t.u./lb.mass/deg.F
d	L	spacing of disturbances in crystal structure
G	FL^{-2}	shear modulus of elasticity
K	$FT^{-1}\theta^{-1}/J$	thermal conductivity, for example, B.t.u./sq.ft./deg.F/hr. for a unit thickness.
k	none	slope of curve of yield point in shear *vs.* normal stress on the shear plane
r_N	L	nose radius of tool
T_0	T	time required for fracture
t	L	depth of cut
V	LT^{-1}	cutting speed – relative speed of tool and workpiece
α	none	rake angle
γ_L	none	ductility—a limiting value of shearing strain
μ	none	a coefficient of rubbing friction—not necessarily simply tan β described previously
ψ_t	θ	a transition temperature of the metal, for example, melting point
ψ_w	θ	workpiece temperature
ρ	ML^{-3}	mass density
τ_0	FL^{-2}	yield stress in shear

However, it is found that the major part of the depth effect is independent of this ratio; a sharp tool taking a shallow cut is not equivalent to a more rounded tool at larger depths but instead will give a much higher specific cutting force. The effect is, therefore, apparently not one that would be observed in a perfectly homogeneous material. A grain size or, possibly better, a mean distance between imperfections of the metal, seems to provide a reasonable choice.

The preceding additional variables might be termed borderline because they apply or become important when the ideal picture starts to break down visibly. While they are very important, it is also true that the experimental results often do not agree well in detail with the present limited theory even when the chip is definitely continuous and the depth of cut is large enough to avoid the equivalent of a size effect. This is not surprising because no mention has been made so far of temperature changes, which are always large locally and usually have an appreciable influence throughout. The tangential force between tool and chip is certainly markedly dependent upon temperature; the yield point in shear will also be affected although the enormously high strain rate may nullify much of the temperature change; ductility will be radically altered; and so on. Again dimensional analysis can point the way.

THE GENERAL CASE

For the analysis of the general case, it is desirable to list all of the independent variables that are to be considered, with their units of force F or mass M, length L, time T, and temperature θ. Both F and M will be used for convenience, $F \equiv MLT^{-2}$. The mechanical equivalent of heat, J, is included to emphasize that a conversion of heat to mechanical units will be necessary at times (see Table I).

If this were a complete list of the independent variables, a functional form of the solution could be obtained easily. There are sixteen quantities and four fundamental units so that twelve independent dimensionless groups exist. If, for example, the shear angle φ, the horizontal force per unit width of cut H, or the temperature rise at the shear plane ψ, is being investigated, then φ, $H/t\tau_0$, or $\psi/(\psi_t-\psi_w)$ can be expressed as a function of $[\alpha, \mu, \gamma_L, k, VT_0/t, r_N/t, d/t, \tau_0/G, \rho V^2/\tau_0, K/Vc\rho t, \tau_0/Jc\rho(\psi_t-\psi_w), \psi_t/\psi_w]$.

The groupings are arbitrary within a very wide choice. Some of the dimensionless variables have been discussed already, the others were chosen because physical meaning could be attached to them easily. The variables α and μ which appear in the idealized theory are definitely important, and k will always have influence. Borderline variables VT_0/t, d/t, and γ_L become important as the continuity of the chip decreases. The ratio r_N/t can usually be ignored for a properly sharpened tool.

More analysis is required for the temperature variables. The shear strain (a function of the independent variables) multiplied by $\tau_0/Jc\rho$ would be the temperature rise at the shear plane if k were zero and if the heat generated were convected but not otherwise conducted away by the metal. $K/Vc\rho t$ is essentially the ratio of the heat conducted from the shear plane to the heat convected by the moving metal. If the temperature rise at the shear plane should be small compared with all significant temperature rises, $\psi_t-\psi_w$, for the metal being cut, then the three-temperature variables can be neglected.

The two remaining groups would seem to be borderline variables, but controlled experimentation is necessary before a definite statement can be made. τ_0/G is the maximum elastic shear strain (the resilience could be used in a somewhat more complicated group). $\rho V^2/\tau_0$ is essentially the ratio of inertia stress to the yield stress in shear and is always extremely small.

So far, the positive aspects of dimensional analysis have been emphasized. Unfortunately, it, too, has its drawbacks. The list of variables may not be sufficiently complete, nor, what is far worse, truly independent. The matter of completeness is rarely obvious, significant omissions may show up when an experimental program cannot be correlated with the use of variables already written down. However, this same trouble must be cleared up before an analytical approach can be successful.

Some of the difficulties can be seen by a re-examination of the list. Quantities like specific heat, thermal conductivity, ductility, coefficient of friction, and yield stress vary not only from one material to the other, but, unpleasantly, also in a given material as the cutting conditions vary. The symbols must therefore be interpreted as the (fixed) values for a given material under a specified set of conditions. If these initial values

and the remaining variables completely determine the values for all other sets of conditions, then the list of variables is complete. This will require at the very least a number of coefficients similar to k, for example, a set of constants for the non-linear variation of specific heat and of thermal conductivity with temperature. If independent quantities of this type exist, no matter how many there may be, there is no essential additional complication as far as an experimental program is concerned although theoretical solutions may be practically ruled out. The metal being cut enters in so many ways that only in the simplest fundamental work will it be possible to substitute other materials, as can very often be done in fluid mechanics problems. Therefore, in most cases, the large number of required constants which should be added to the list of variables will automatically be the same in model and prototype, experiment and production, or two production runs. In fact, some of the dimensionless combinations already considered are similarly fixed. Therefore, at appreciable depths of cut, probably only

$$(\alpha, VT_0/t, K/Vc\rho t)$$

are of major significance *for a given material and temperature of workpiece.*

It is interesting that almost as in the Reynolds and Froude numbers for fluids, V and t appear in ratio in one-dimensionless variable and as a product in the other. Therefore, complete similarity between two experiments cannot be obtained using one material and is rarely practicable. This discouraging result is almost always found when the list of variables is reasonably extensive. However, it may be true, as for fluids, that the range of importance of each dimensionless group containing V, t is markedly separate, or that the phenomena they control are quite distinct. There is a strong probability that this is true for metal cutting. VT_0/t seems to control the degree of continuity of the chip while $K/Vc\rho t$ should be of major importance in the definitely continuous range. Only careful experimental investigations of many different metals will provide the answer.

In the previous discussion, the tool itself has hardly been considered. Here, again, it is likely that in only a few cases can information be obtained with one tool material and applied to another in a quantitative manner. Also, the use of a cutting fluid or coolant would add many variables. All of these difficulties merely emphasize the impossibility at present of developing a mathematical theory which could quantitatively cover even a small section of the complete problem, and demonstrate the necessity of carefully conceived experimental investigations based on dimensional analysis.

CONCLUSION

The best procedure to obtain a clear picture of metal cutting would seem to be an experimental analysis of dry two-dimensional cutting on the basis of the dimensionless variables already discussed. Despite the very large volume of data on the cutting of metals there is only a little information that is directly useful for this particular fundamental purpose.

The two main variables, V and t, appearing in VT_0/t and $K/Vc\rho t$, are of prime significance because they can be varied in a known and simple manner, but most of the others are of equal physical importance and may merit careful study. Those involving workpiece temperature directly can also be varied without too much difficulty and offer interesting possibilities.

It is quite likely that additional variables will be found necessary to explain all results and that a rearrangement of some of the suggested dimensionless groups will be advantageous.

ACKNOWLEDGMENT

The authors wish to thank the Shell Oil Company for permission to publish this paper.

The Current Build-Up in a Planar Diode

EDWARD H. GAMBLE*
Polytechnic Institute of Brooklyn, Brooklyn, New York
(Received June 29, 1949)

A theoretical study has been made of the initial electron flow when a signal voltage $V(t)$ is applied across an idealized planar diode. The analysis is based upon three fundamental postulates: Poisson's equation, Newton's equation, and the continuity equation. This quasi-stationary study includes both the growth and motion of the space-charge cloud of electrons across the interelectrode space and the transient build-up of the current in the external circuit. The analytical forms and numerical plots of the current build-up under space-charge-limited and temperature-limited operation are compared for various "large" and "small" signal voltages.

STATEMENT OF THE PROBLEM

A STUDY of the initial electron flow when a signal voltage $V(t)$ is applied across a planar diode has been undertaken. This problem may be attacked from two possible directions. One can make a study of the way in which the space-charge cloud of electrons forms and determine the motion of this charge under the action of the electric field due to the applied signal and the counter effect of the charge upon this field. The other approach is that of studying the transient build-up of the current in the external circuit.

This latter approach is the one which will be adopted to secure the immediate results of interest to the electronics engineer. By analogy to hydrodynamics, one may specify the quantities describing the moving electron either in the Lagrange variables (t,τ) or the Euler variables $[t,x]$. t is the clock time measured from the instant when the voltage is applied to the diode, and τ denotes the instant when a particular group of electrons is emitted from the cathode surface. τ is referred to as the "emission" time.

The diode configuration is shown in Fig. 1.

DETERMINATION OF THE DENSITY OF THE SPACE-CHARGE CLOUD

The symbol Φ denotes the scalar potential, E the electric field density, and ρ the space-charge density.

If one makes the assumption that any signal voltages which one may apply across the electrodes of the vacuum diode do not change significantly in the time required for an electromagnetic wave to propagate the interelectrode space, one may treat the entire problem as quasi-stationary. The three basic electronic equations reduce to Poisson's equation, Euler or Newton's equation of motion, and the continuity equation. The formulation of the important physical quantities then becomes

$$E[t,x]=-\frac{V(t)}{h}+\frac{1}{h\epsilon_0}\int_0^h(h-\xi)\rho[t,\xi]d\xi-\frac{1}{\epsilon_0}\int_0^x\rho[t,\xi]d\xi \quad (1)$$

$$\Phi[t,x]=\frac{x}{h}V(t)-\frac{x}{h\epsilon_0}\int_0^h(h-\xi)\rho[t,\xi]d\xi+\frac{1}{\epsilon_0}\int_0^x(x-\xi)\rho[t,\xi]d\xi \quad (2)$$

$$I(t)=\frac{\epsilon_0}{h}\frac{d}{dt}V(t)+\frac{1}{h}\int_0^h\left\{-\rho[t,x]\frac{\partial x}{\partial t}\right\}dx \quad \text{and}$$

$$=\frac{\epsilon_0}{h}\dot{V}(t)+I_{\text{ind}}(t), \quad (3)$$

where the subscript *ind* signifies that the quantity is related to the current induced in the external circuit due to the motion of the electrons.

$$x(t,\tau)=u(\tau,\tau)[t-\tau]+\frac{e}{mh}\int_\tau^t(t-\eta)V(\eta)d\eta-\frac{e}{mh\epsilon_0}\int_\tau^t(t-\eta)\left\{\int_0^h(h-\xi)\rho[\eta,\xi]d\xi\right\}d\eta+\frac{e}{m\epsilon_0}\int_\tau^t(t-\eta)\left\{\int_0^x\rho[\eta,\xi]d\xi\right\}d\eta, \quad (4)$$

where $u(\tau,\tau)$ denotes the initial velocity of the τ-group of electrons.

If one restricts the problem to the initial electron flow and hence the period of time to that of the transit time of the first electron,

$$0<t<t_1, \quad \tau=0, \quad (5)$$

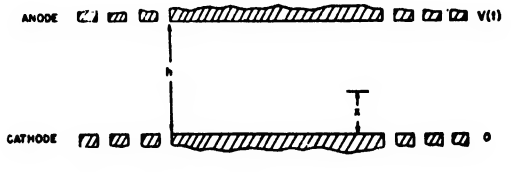


FIG. 1. Planar diode configuration.

* Now with Curtiss-Wright Corporation, Columbus, Ohio.

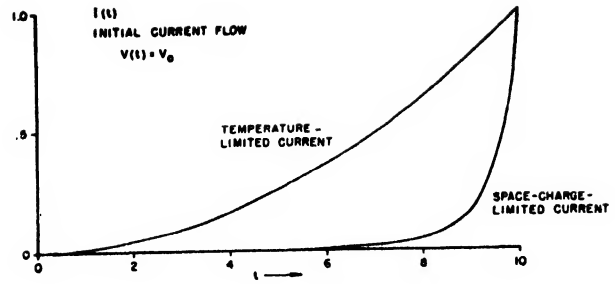


FIG. 2.

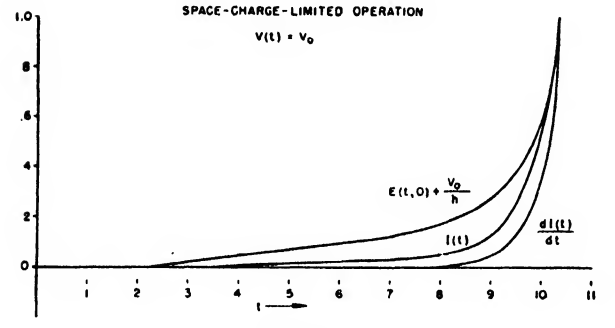


FIG. 3.

Eq. (4) becomes

$$x(t,0)=u(0,0)t+\frac{e}{mh}\int_0^t(t-\eta)V(\eta)d\eta$$
$$-\frac{e}{mh\epsilon_0}\int_0^t(t-\eta)\left\{\int_0^h(h-\xi)\rho[\eta,\xi]d\xi\right\}d\eta$$
$$+\frac{e}{m\epsilon_0}\int_0^t(t-\eta)\left\{\int_0^x\rho[\eta,\xi]d\xi\right\}d\eta. \quad (6)$$

Equation (6) may be considered as an integral equation for $\rho[t,x]$ when one specifies the position x and the time t. Knowing the form of $\rho[t,x]$, this quantity may be substituted into Eqs. (1)–(3) to determine the other desired quantities.

DETERMINATION OF THE EXTERNAL CURRENT

If one chooses the total current $I(t)$ as the unknown quantity to be found, the important equations are

$$E(t,\tau)=E_c(\tau,\tau)-\frac{V(t)}{h}-\frac{1}{\epsilon_0}\{q_{\text{ind}}(t)-q_{\text{ind}}(\tau)\}$$
$$=E_c(\tau,\tau)-\frac{V(t)}{h}-\frac{1}{\epsilon_0}\int_\tau^t I_{\text{ind}}(\eta)d\eta, \quad (7)$$

where $E_c(\tau,\tau)$ is the field at the cathode at the instant τ, and

$$q_{\text{ind}}(t)=\int_0^t I_{\text{ind}}(\eta)d\eta, \quad (8)$$

$$\Phi(t,\tau)=-\int_\tau^t E(t,\eta)\frac{\partial x(t,\eta)}{\partial\eta}d\eta$$
$$=-E_c(\tau,\tau)x(t,\tau)+\frac{x(t,\tau)}{\epsilon_0}\int_\tau^t I(\eta)d\eta$$
$$+\int_\tau^t x(t,\eta)\frac{\partial E(\eta)}{\partial\eta}d\eta, \quad (9)$$

$$x(t,\tau)=u(\tau,\tau)[t-\tau]-\frac{e}{2m}E_c(\tau,\tau)[t-\tau]^2$$
$$+\frac{e}{m\epsilon_0}\int_\tau^t\frac{(t-\eta)^2}{2}I(\eta)d\eta$$
$$=u(\tau,\tau)[t-\tau]-\frac{e}{2m}E_c(\tau,\tau)[t-\tau]^2$$
$$+\frac{e}{mh}\int_\tau^t(t-\eta)V(\eta)d\eta$$
$$+\frac{e}{m\epsilon_0}\int_\tau^t\frac{(t-\eta)^2}{2}I_{\text{ind}}(\eta)d\eta. \quad (10)$$

SPACE-CHARGE-LIMITED CURRENT BUILD-UP

If one applies the boundary conditions

$$E_c(\tau,\tau)=E[t,0]=0, \quad u(\tau,\tau)=0, \quad (11)$$

the relations given above reduce to those found previously by Wang[1] for the space-charge-limited planar diode. For conditions (5), the basic integral equation becomes

$$V(t)=\frac{h}{\epsilon_0}q(t)-\frac{e}{m\epsilon_0^2}\int_0^t\frac{(t-\eta)^2}{2}q(\eta)I(\eta)d\eta. \quad (12)$$

Differentiating Eq. (12) twice with respect to t, one finds that

$$\frac{\partial^2V}{\partial t^2}=\frac{h}{\epsilon_0}\ddot{q}(t)+\frac{e}{2mh}q^2(t). \quad (13)$$

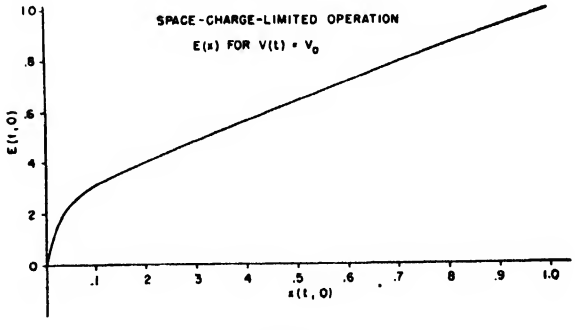


FIG. 4.

[1] C. C. Wang, "Large signal high frequency electronics of thermionic vacuum tubes," Proc. I. R. E. (April, 1941).

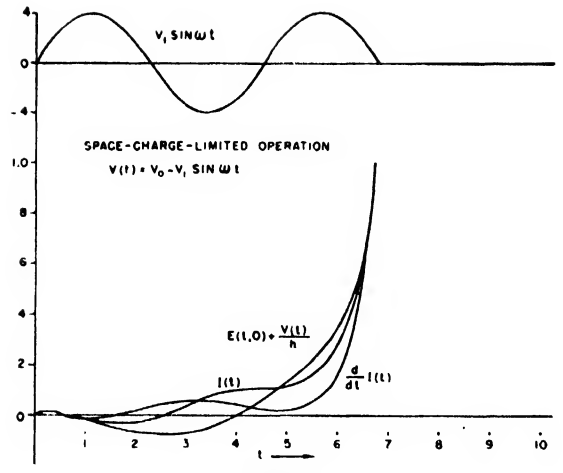


FIG. 5.

The solution of Eq. (13) for $(d/dt)q(t)=I(t)$ will give the current build-up for the space-charge-limited diode during this initial period before the electrons strike the anode.

TEMPERATURE-LIMITED CURRENT BUILD-UP

The conditions for temperature-limited operation may be defined as

$$I_0=\text{saturation current}=\left\{-\rho[t,x]\frac{\partial x}{\partial t}\right\} \tag{14}$$

$$E[t,x]=-V(t)/h=E(t) \tag{15}$$

$$\Phi[t,x]=[x(t,\tau)/h]V(t). \tag{16}$$

For the period before any electrons strike the anode,

$$x(t,\tau)=u(\tau,\tau)[t-\tau]+\frac{e}{mh}\int_\tau^t(t-\eta)V(\eta)d\eta \tag{17}$$

$$\Phi(t,\tau)=\frac{u(\tau,\tau)}{h}[t-\tau]V(t)+\frac{V(t)}{mh^2}\int_\tau^t(t-\eta)V(\eta)d\eta \tag{18}$$

$$I(t)=\frac{\epsilon_0}{h}\dot{V}(t)+\frac{I_0}{h}u(\tau,\tau)[t-\tau]+\frac{I_0e}{mh^2}\int_\tau^t(t-\eta)V(\eta)d\eta. \tag{19}$$

SOME SPECIFIC CASES

We now have all of the basic formulas necessary to determine the initial current build-up. It will be advantageous to compare the space-charge-limited and the temperature-limited cases. These are the solutions of Eqs. (13) and (19), respectively, for conditions (5).

Let $V(t)=V_0$, so that the build-up for a constant applied signal with its application to pulses may be determined. Figure 2 shows an enlarged plot of the solutions for this case. This plot shows the final current rising to the same value for both classes of operation. This will, in general, not be true. The functional form of the current as a function of time is of importance.

The temperature-limited current according to Eq. (19) is

$$I_{\text{ind}}(t)=(I_0e/mh^2)V_0(t^2/2) \tag{20}$$

for negligible initial velocity $u(\tau,\tau)$.

The position of the leading edge of the electron cloud is given by Eq. (17).

$$x(t,0)=u(\tau,\tau)\cdot t+(V_0/2mh)t^2. \tag{21}$$

The current and space charge always build up as the same power of time t for the temperature-limited operation.[2]

The numerical solution of Eq. (13) indicates that the space-charge-limited current rises much more rapidly than a square of the clock time. An accurate empirical formula for this current is given by Eq. (22)

$$I_{\text{ind}}(t)=I_{\text{const}}\left\{30.121\left(\frac{t}{t_1}\right)^{10}-51.009\left(\frac{t}{t_1}\right)^{9}+21.889\left(\frac{t}{t_1}\right)^{8}\right\}. \tag{22}$$

The complete analytical solution of Eq. (13) for this case is an elliptic integral

$$I_{\text{ind}}(t)=\dot{q}_{\text{ind}}(t)=\int_0^t\left[\int_0^\eta I_{\text{ind}}(\xi)d\xi\right]^2 d\eta. \tag{23}$$

For this case, Eq. (7) reduces to

$$E(t,0)+\frac{V_0}{h}=-\frac{q_{\text{ind}}(t)}{\epsilon_0}=-\frac{1}{\epsilon_0}\int_0^t I(\eta)d\eta$$

$$=-\frac{1}{\epsilon_0}I_{\text{const}}\left\{39.405\left(\frac{t}{t_1}\right)^{11}-73.406\left(\frac{t}{t_1}\right)^{10}+35.001\left(\frac{t}{t_1}\right)^{9}\right\}. \tag{24}$$

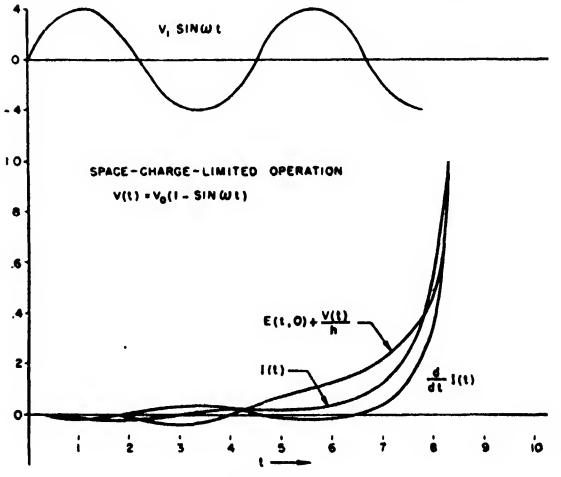


FIG. 6.

[2] B. Salzberg, "Large signal high frequency theory of the vacuum tube," dissertation for D.E.E. at the Polytechnic Institute of Brooklyn, Brooklyn, New York (1941).

Hence, a plot of $q(t)$, as shown in Fig. 3, is also a plot of the electric field as a function of time. For the determination of the leading edge of the electron cloud, one must employ Eq. (10)

$$x(t,0)=\frac{e}{2m\epsilon_0}30.121I_{\text{const}}\int_{\tau=0}^{t}(t-\eta)^2F(\eta)d\eta, \quad (25)$$

where

$$F(\eta)=\left[\frac{(t-\eta)^{10}}{t_1^{10}}-1.6935\frac{(t-\eta)^9}{t_1^9}+0.7267\frac{(t-\eta)^8}{t_1^8}\right]$$

$$x(t,0)=\frac{e}{m\epsilon_0}15.06I_{\text{const}}\left[\left(\frac{t}{t_1}\right)^{10}-1.6935\left(\frac{t}{t_1}\right)^9-0.7267\left(\frac{t}{t_1}\right)^8\right]t^2. \quad (26)$$

A plot can be made of $E(x)$ by a comparison of $E(t,0)$ given by the uppermost plot of Fig. 3 and the data computed from Eq. (26). This is shown in Fig. 4.

For the applied signal

$$V(t)=V_0-V_1\sin\omega t \quad (27)$$

for

$$V_1>V_0, \quad \omega>1/t_1, \quad (28)$$

the functional forms of $I(t)$ and $(d/dt)I(t)$ are shown in Fig. 5. The function

$$[E(t,0)+(V(t)/h)]$$

is also plotted in Fig. 5. The electric field $E(t,0)$ for the applied voltage given by Eq. (27) is

$$E(t,0)=-\frac{V_0}{h}+\frac{V_1\sin\omega t}{h}-\frac{1}{\epsilon_0}\int_{\tau=0}^{t}I_{\text{ind}}(\eta)d\eta. \quad (29)$$

The actual value of the ratio V_0/V_1 is not determinable from the numerical solution of Eq. (13) since only the second time derivative of the applied signal

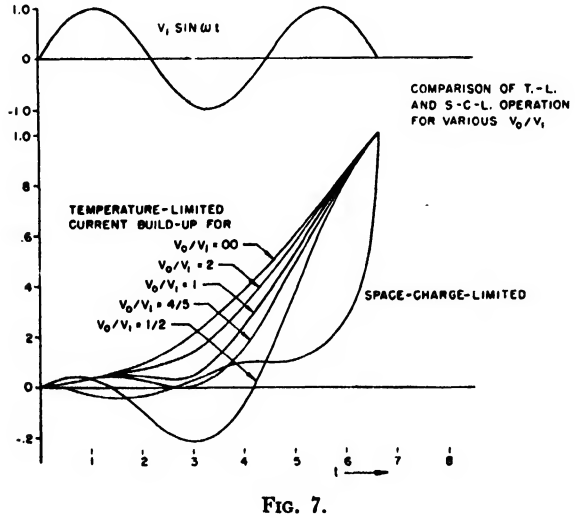


FIG. 7.

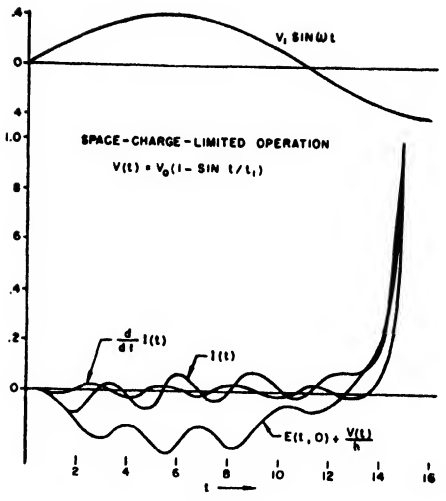


FIG. 8.

enters into the equation. However, experience has shown that if this ratio becomes too much smaller than *one*, the field will be negative and the electrons will never reach the anode. The period of the applied signal is clearly shown on the graph in order that one may visualize the time sequence. For the particular case where $V_0/V_1=1$, Fig. 6 shows the form of the current and the electric field as functions of time. A comparison of Figs. 5 and 6 for $I(t)$ indicates that the period of time during which the current is constant has been lengthened and the build-up time extended when the full modulation effect takes place for the amplitude ratio *one*.

It was considered advisable to make a direct comparison of the current build-up for space-charge-limited operation and temperature-limited operation for the applied signal given by Eq. (27) and conditions (28). Figure 7 shows this comparison. Several values of the ratio V_0/V_1 were chosen and plots made. The formula for the temperature-limited current is found by the substitution of Eq. (27) into Eq. (19).

$$I(t)=V_1\left\{-\frac{\epsilon_0}{h}\omega\cos\omega t+\frac{I_0e}{mh^2}\left[\frac{V_0}{V_1}\frac{t^2}{2}-\frac{t}{\omega^2}+\frac{\sin\omega t}{\omega^2}\right]\right\}. \quad (30)$$

The uppermost plot of Fig. 7 shows directly the comparison of the space-charge-limited current and the temperature-limited current for the case $V_0/V_1=\frac{1}{2}$. Another interesting plot is given in Fig. 7 which shows the variation of the build-up current with changes of the ratio V_0/V_1. This covers most of the range of large signals.

When the frequency of the time-varying component of the signal voltage is nearly equal to the reciprocal of the transit time, one finds that the build-up time is lengthened since the electric field is negative for a longer period of time. One point of interest is the large difference in the period of oscillation of the electric field and the current and the period of the signal com-

ponent. This is clearly shown in Fig. 8. The basic form of the signal voltage is

$$V(t)=V_0(1-\sin t/t_1). \quad (31)$$

CONCLUSIONS

The initial current build-up of thermionic vacuum tubes under space-charge-limited operation is a function of higher powers of time than under temperature-limited operation. The build-up may be considerably delayed or shortened when sinusoidally varying signals whose amplitude is comparable to the constant signal voltage are applied. The results shown in this paper should be of particular interest to engineers studying microsecond pulsing of thermionic vacuum tubes.

ACKNOWLEDGMENT

The author wishes to express his gratitude to the ONR who sponsored a special fellowship which enabled the author to carry on the research work given in this paper. A vote of thanks is due to the Reeves Instrument Company which made their analog computer available for numerical solutions. The author shall always be indebted to Dr. Ernst Weber of the Microwave Research Institute whose kindly advice and supervision served as a driving force and inspiration.

Thermocouples of the Refractory Metals

F. Haydn Morgan and W. E. Danforth

Bartol Research Foundation of The Franklin Institute, Swarthmore, Pennsylvania

(Received July 13, 1949)

The thermoelectric power of several refractory thermocouples was investigated, W-Ta for temperature measurements up to 3000°C, W-Mo up to 2600°C, Ta-Mo up to 2600°C and W-W/Mo (alloy 50 percent Mo/50 percent W) up to 2900°C.

Basic calibration was carried out using several melting points and the results were verified and interpolated by means of an optical pyrometer.

The couples W-Ta and W-Mo have been applied in this laboratory to the temperature measurement requirements in high temperature vacuum furnaces and determinations of spectral emissivity.

IN order to satisfy the need for accurate measurements of high temperature in electronic research, the thermocouples of the refractory metals have been investigated. Existing literature did not exceed 2000°C, with the exception of Morugina,[1] who studied W-Ta; and Pirani and Wagenheim,[2] who reported on W-W/Mo (75 percent W/25 percent Mo).

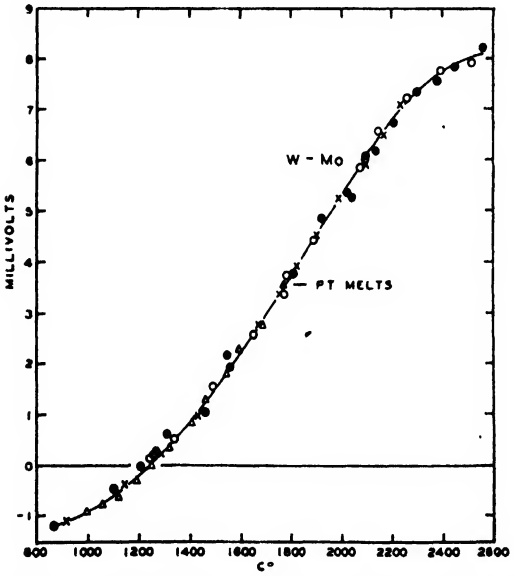


FIG. 1. Composite calibration data for W-Mo thermocouple. The different types of points represent different runs with optical pyrometer. The solid triangle represents melting point data taken with platinum in a vacuum furnace.

For the optical pyrometer comparisons the thermocouple wires 0.005 in. or 0.002 in. diameter were spot-welded to a 0.030-in. tungsten wire, which was heated by passage of current. This was placed in a bell jar and evacuated to 10^{-5} mm Hg. A Leeds and Northrup Type K potentiometer was used to measure the e.m.f. and a Leeds and Northrup Optical Pyrometer to measure the brightness temperature. Using 0.43 as the spectral emissivity[3] the brightness temperature was converted to true temperature. The cold junction was held at room temperature.

The melting point calibrations were obtained by spot-welding the thermocouple to the midpoint of a 2-cm length of 0.030-in. wire of the standard metal. The following metals were used:[4] Gold 1063.0°C ±0.0°, Nickel 1455°C±1°, Platinum 1773.5°C±1°, Molybdenum 2625°C±50°, and Tantalum 3000°C ±100°. An optical pyrometer would detect no temperature differences in the vicinity of the couple. The po-

[1] S. Morugina, Zeits. f. Tech. Physik **7**, 486 (1926).

[2] M. Pirani and G. Wagenheim, Zeits. f. Tech. Physik **6**, 358 (1925).

[3] W. F. Roeser and H. T. Wensel, *Handbook of Chemistry and Physics* (1947), thirtieth edition, p. 2246.

[4] National Bureau of Standards Miscellaneous Publication M183.

TABLE I. E.m.f. (millivolts).

T(°C)	Ta-Mo	W-Ta	W-Mo	W-W/Mo
800	12.20	10.65	−1.32	−0.50
900	13.60	12.20	−1.15	−0.30
1000	14.80	13.72	−0.80	0.00
1063		14.70		
1100	15.95	15.25	−0.60	0.30
1200	16.90	16.65	−0.20	0.70
1250	17.32	17.32	0.00	
1300	17.70	17.95	0.30	1.15
1400	18.35	19.05	0.85	1.65
1455		19.65		
1500	18.80	20.05	1.50	2.25
1600	19.25	20.90	2.18	2.85
1700	19.45	21.65	2.95	3.45
1773		22.12	3.54	
1800	19.45	22.25	3.73	4.00
1900	19.25	22.70	4.50	4.60
2000	18.85	22.90	5.30	5.20
2100	18.35		6.05	5.80
2200	17.70		6.80	6.40
2300	16.90		7.35	6.90
2400	15.90		7.75	7.35
2500	14.90		8.00	7.70
2600				7.95
2625		20.90		
2700				8.08
2800				8.12
2900				8.18
3000		18.50		

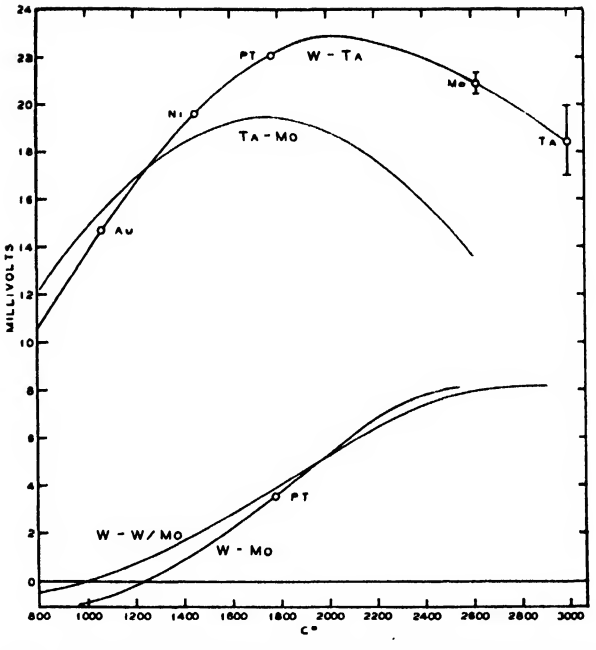


FIG. 2. Calibration curves of four refractory thermocouples. Melting point data are indicated on the W-Ta and W-Mo curves.

tentiometer was kept in balance as the temperature was slowly raised. When the wire melted, permitting the thermocouple to cool, the potentiometer would suddenly go out of balance. The last point of balance was taken as the melting point.

Considerable time was spent verifying the reproducibility of the present results. Wire stock from different spools and of different sizes was tested. Figure 1 shows composite calibration results in the case of the W-Mo couple. Similar sets of data were taken with the other couples and the resulting average curves are plotted in Fig. 2 and presented in Table I.

W-Ta (contrary to Morugina's findings of a straight line) gives a curve which rises from 14 millivolts at 1000°C to a maximum of 22.9 millivolts at 2000°C and then decreases as the temperature is increased. The slope of the curve between 1300°C and 1800°C is 8 microvolts per degree. The couple was found to be reproducible up to 2000°C, for which the graph shows pyrometer data. Above this temperature the pyrometer data showed excessive spread and, for some reason that was not determined, did not agree with the melting points of Mo and Ta which have been plotted.

The W-Mo curve has negative values below 1250°C, where it passes through zero and then rises to 8 millivolts at 2500°C. The average slope of the curve between these temperatures is 6.5 microvolts per degree. The scatter of the data in this case was relatively small, and agrees with literature published by B. Osann and Schroeder,[5] which gave the W-Mo curve up to 1850°C.

The W-W/Mo (50 percent W–50 percent M) curve is of very similar form to the W-Mo curve. It passes through zero at 1000°C and rises to 8 millivolts at 2600°C where it tends to saturate. The slope of the curve is 5 microvolts per degree.

The Ta-Mo curve crosses the W-Ta curve at 1250°C or 17.3 millivolts. This temperature value agrees, as it should, with the temperature for zero output of the W-Mo couple. It reaches a maximum of 20.2 millivolts at 1700°C. Due to this maximum at such a low temperature, it has not been used for practical applications.

The W-Ta couple is being used for spectral emissivity determination of different materials. The couple is spot-welded to a 0.030-in. tungsten wire which is coated with the material being investigated. The W-Mo couple is being used for temperature measurements in a vacuum furnace, where optical pyrometry is difficult.

[5] B. Osann and E. Schroeder, Arch. f. d. Eisenhuttenwesen 7, 89 94 (1933).

Interaction of a Spiral Electron Beam and a Resonant Microwave Cavity*

F. K. Willenbrock and S. P. Cooke
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received July 18, 1949)

The interaction between the electromagnetic field in a cavity and an electron beam projected along the axis of the cavity is examined. The particular cavity considered here is of the cylindrical TE11n_s type in a steady axial magnetic field. If the cavity is excited in a linearly polarized mode, the electromagnetic field will drive the electrons in a helical trajectory with an expanding radius, and the electrons will excite and transfer energy to a degenerate mode oriented spatially at right angles to the driving field. In the driving plane of polarization (both planes of polarization if the cavity is excited in a circularly polarized mode), the electron beam will excite a field in phase opposition to the driving field in a manner analogous to the counter e.m.f. in an electromechanical generator. The converse case of a TE11n_s cavity excited by a spiral beam of electrons is also considered.

I. QUALITATIVE DISCUSSION OF SPIRALING ELECTRON BEAMS

THE behavior of a beam of electrons in a crossed electric and magnetic field has been widely discussed. In the recent literature,[1–3] several different uses of spiraling beams have been suggested. Malter[1] examines the impedance change in a parallel plate condenser through which an electron beam is projected in the presence of a steady magnetic field. Smith and Schulman[2] discuss the effect of spiraling electron beams on the frequency and Q of resonant systems. Spiral beam control has been used for frequency and amplitude modulation in several experimental magnetrons.[4–6]

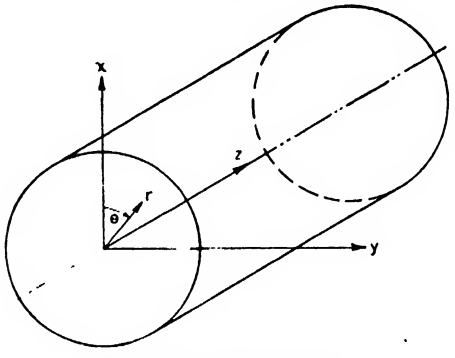


Fig. 1. Coordinate system.

Truell[3] considers the possibility of measuring the amplitude of an electromagnetic field by the deflection of an electron beam in the presence of a steady magnetic field.

The present investigation is a study of the interaction of an axial beam of electrons with a cylindrical cavity oscillating in a TE11n_s mode with a steady magnetic field superimposed along the axis of the cavity. The theory of the interaction is presented both from the standpoint of energy transfer from the electromagnetic field to the electron beam and from the beam to the field.

To simplify the analysis, the trajectory of a single electron is considered first. A single spiraling electron, unlike an electron beam, does not cause a significant change of the r-f field in the cavity, and does not extract an appreciable part of the power driving the cavity. The electron follows a helical trajectory with an increasing radius in a cavity driven by a linearly or by a circularly polarized field. Linear polarization is obtained when the cavity is driven through a single coupling iris; circular polarization is obtained when the cavity is driven through two coupling holes located at right angles in space with equal input-signal amplitudes in time quadrature.

For a given cavity, the two parameters that can be readily varied are the magnetic field strength which determines the cyclotron angular frequency, $\omega_c = eB_0/m$, and the axial velocity of the electron beam, v_0, which determines the transit time of the electron through the cavity. It is found that for a cavity of length $\lambda_g/2$, maximum energy is transferred to the electron when $\omega_c = \omega_0$, the resonant angular frequency of the cavity. For a cavity of length λ_g with $\omega_0 = \omega_c$, all or nearly all the energy gained by the beam during the first half-wave-length will be returned to the r-f field during the second half-wave-length; and there will be a negligible amount of rotational energy in the beam at exit. The deflection of the electron varies inversely with v_0.

The treatment of the trajectory of a single electron in a cavity excited in a linearly polarized mode is also valid for an electron beam if (1) the effect of space charge is neglected; (2) the degenerate mode at right angles to the driving mode is suppressed; (3) the de-

* This material was adapted from Cruft Laboratory Technical Report No. 79 entitled, "Modulation, generation, and frequency multiplication of microwave oscillations by a spiral beam of electrons. I." The research was made possible through support extended Cruft Laboratory, Harvard University jointly by the Navy Department (ONR), the Signal Corps of the U. S. Army, and the U. S. Air Force, under Contract N5ori-76, T.O. 1.

[1] L. Malter, "Deflection and impedance of electron beams at high frequencies in the presence of a magnetic field," RCA Rev. 5, 439–454 (April, 1941).

[2] L. P. Smith, and C. I. Schulman, "Frequency modulation and control of electron beams," Proc. I.R.E. 35, 644–657 (July, 1947).

[3] R. Truell, "A method of measuring the field strength of high-frequency electromagnetic fields," Proc. I.R.E. 36, 1249–1251 (October, 1948).

[4] Kilgore, Schulman, and Kurshan, "A frequency-modulated magnetron for super-high frequencies," Proc. I.R.E. 35, 657–664 (July, 1947).

[5] Donal, Bush, Cuccia, and Hegbar, "A 1-Kilowatt Frequency-Modulated Magnetron for 900 Megacycles," Proc. I.R.E. 35, 664–669 (July, 1947).

[6] J. S. Donal, Jr. and R. R. Bush, "A spiral-beam method for the amplitude modulation of magnetrons," Proc. I.R.E. 37, 375–382 (April, 1949).

crease of the effective driving field to compensate for the energy transferred to the electron beam is taken into account.

The effect of space charge can be neglected because of the relatively low charge densities in the beam and the presence of the steady axial magnetic field. The beam in the cavity is not bunched as in a magnetron, but is a steady stream of electrons. Although the cross section of the beam is assumed to be small, the total beam currents considered are also small. For these reasons, high charge densities are not obtained. In any event, the presence of the strong axial magnetic field prevents defocusing of the beam.

Unless the perpendicular mode which is degenerate with the driving mode is suppressed, it will be excited by the spiraling electron beam. The r-f field in the cavity is, therefore, a superposition of a linear driving mode and a mode at right angles driven by the electron beam. This results in an elliptically polarized r-f field. If this degenerate mode is suppressed, the trajectory of the beam is the same as that of a single electron provided the decrease of effective driving field is considered.

The electron beam also excites a mode of oscillation in the original plane of polarization in phase opposition to the driving field. This "counter" E-field reduces the effective amplitude of the driving field and is analogous to the counter-e.m.f. in an electromechanical generator. With this reduced driving field, the losses in the cavity walls are also decreased. The remaining r-f input power is transferred by the beam to the other plane of polarization and appears as rotational energy in the beam at exit.

For the circularly polarized mode, there is a field driven through an iris by the external generator in each plane of polarization, and in each of these planes there are two counter-fields in phase opposition to the externally driven field. The first of these counter-fields is the same as that described for the linear case and is analogous to the counter-e.m.f. in an electromechanical generator. This is the "direct" counter-field which would exist even if the mode in the perpendicular plane were suppressed.

The second counter-field is cross-coupled by the electron beam from the driving field in the perpendicular plane of polarization. In the linear case this cross-coupled field drives the other degenerate mode and produces elliptical polarization. In the circular case, this cross-coupled field appears in opposition to the phase of the externally driven field if the latter has been chosen to add energy to the beam. The two counter-fields reduce the effective field more than the single counter-field in the linear case, and there is a correspondingly greater energy transfer to the beam for circular polarization.

If a spiraling beam of electrons is projected along the axis of the second TE_{11n_z} cavity, the cavity is excited, and energy is transferred from the beam to the r-f field. With the proper choice of parameters, it is possible to extract all the rotational energy from the beam. However, if the transit time is too long, the electron beam regains rotational energy from the r-f field in the cavity.

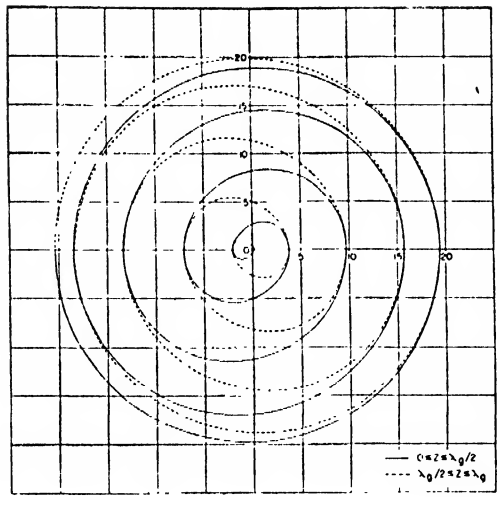


FIG. 2. Projection on the transverse plane of the trajectory of a single electron. TE_{11} cavity excited in a linearly polarized mode. $\kappa=1$, $\eta=0.1$, and $\phi=0$.

II. TRAJECTORY OF AN ELECTRON IN A TE_{11n_z} CAVITY IN A MAGNETIC FIELD

The transverse E-field in a perfectly conducting TE_{11n_z} cavity can be represented as $\mathbf{E}(r,\theta)\sin 2\pi z/\lambda_g \cos\omega_0 t$. For small deviations from the axis of the cavity $\mathbf{E}(r,\theta)$ can be written, by proper choice of the x axis, as $E_x=E_{0x}$, $E_y=0$, where E_{0x} is a constant. For velocities small in comparison with the velocity of light and for small r-f power inputs, the effect of the r-f magnetic field is negligible. The nonrelativistic equations of motion can be written† (see Fig. 1),

$$\left.\begin{aligned} \ddot{x} &= -e\mathcal{E}_x/m-(e/m)B_0\dot{y} \quad \text{(a)} \\ \ddot{y} &= -e\mathcal{E}_y/m+(e/m)B_0\dot{x} \quad \text{(b)} \\ \ddot{z} &= 0 \quad \text{(c)} \end{aligned}\right\} \tag{1}$$

where the dot notation indicates derivatives with respect to time, and e/m is the charge-to-mass ratio of the electron. The steady magnetic field, B_0, is directed along the positive z axis.

A linearly polarized mode as seen by the electron in its transit through the cavity can be written as

$$\left.\begin{aligned} \mathcal{E}_x &= E_{0x}\sin(2\pi z/\lambda_g)\cos(\omega_0 t+\phi) \\ \mathcal{E}_y &= \mathcal{E}_z = 0 \end{aligned}\right\} \tag{2}$$

for the interval $0\leq t\leq t_1$. Here, λ_g = wave-length in the cavity, $f_0=\omega_0/2\pi$ = resonant frequency of the cavity, t_1 = transit time of the electron through the cavity, and ϕ = phase angle of E-field at time of entrance.

† Rationalized M.K.S. units are used in this report.

The solution of Eq. (1c) is simply $z=v_0t$ where v_0 is the longitudinal velocity of the electron, and $z=0$ at $t=0$. Substituting this expression in the above equations will lead to solutions for $x(t)$ and $y(t)$ in terms of the parameters $\kappa=\omega_0/\omega_c$ and $\eta=v_0/v_p$, where $\omega_c=(e/m)B_0$ is the cylotron angular frequency, and v_p is the phase velocity of the r-f wave in the cavity. The solutions are for $\kappa=1$ (see Appendix A),

$$\left.\begin{aligned} x(t)=&-\frac{eE_{0x}}{2m\omega_0^2}\left\{\frac{\sin[(1+\eta)\omega_0t+\phi]}{\eta(2+\eta)}\right.\\ &-\frac{\sin[(1-\eta)\omega_0t+\phi]}{\eta(2-\eta)}+\frac{2}{\eta(4-\eta^2)}\\ &\left.\times[(2-\eta^2)\sin\omega_0t\cos\phi+2\cos\omega_0t\sin\phi]\right\}\\ y(t)=&-\frac{eE_{0x}}{2m\omega_0^2}\left\{-\frac{2\eta\cos}{1-\eta^2}+\frac{\cos[(1+\eta)\omega_0t+\phi]}{\eta(2+\eta)(1+\eta)}\right.\\ &+\frac{\cos[(1-\eta)\omega_0t+\phi]}{\eta(2-\eta)(1-\eta)}-\frac{2}{\eta(4-\eta^2)}\\ &\left.\times[(2-\eta^2)\cos\omega_0t\cos\phi-2\sin\omega_0t\sin\phi]\right\}\end{aligned}\right\}. \quad (3)$$

The electron traces a trajectory in the form of a helix with an expanding radius. Figure 2 shows a projection on the transverse (x, y) plane of the trajectory of the electron as it travels longitudinally through the cavity. For the chosen direction of the B-field, the electron spirals in a clockwise sense. The axes of the figures are

$$x'(t)=x(t)\Big/-\frac{eE_{0x}}{2m\omega_0^2}$$

and

$$y'(t)=y(t)\Big/-\frac{eE_{0x}}{2m\omega_0^2}.$$

This curve is plotted for $\eta=0.1$, which corresponds approximately to a 4000-volt electron. This high value of voltage was chosen to facilitate the calculation and plotting. For a more efficient transfer of energy to the beam, accelerating voltages of the order of 100 volts or less should be used.

Similarly, a circularly polarized mode as seen by the electron in the cavity can be written as

$$\left.\begin{aligned}\mathcal{E}_x&=E_{0x}\sin(2\pi/\lambda_g)z\cos\omega_0t\\ \mathcal{E}_y&=E_{0x}\sin(2\pi/\lambda_g)z\sin\omega_0t\\ \mathcal{E}_z&=0\end{aligned}\right\} \quad (4)$$

for the interval $0\leq t\leq t_1$. The direction of the B-field is taken the same as in the linearly polarized case. The circularly polarized field is chosen so that the E-field vector rotates in the clockwise sense. The solutions of the equations of motion for the same conditions are (see Appendix A),

$$\left.\begin{aligned} x(t)=&-\frac{eE_{0x}}{2m\omega_0^2}\left\{\frac{2}{\eta}\sin\omega_0t-\frac{1}{\eta(1+\eta)}\sin(1+\eta)\omega_0t\right.\\ &\left.-\frac{1}{\eta(1-\eta)}\sin(1-\eta)\omega_0t\right\}\\ y(t)=&-\frac{eE_{0x}}{2m\omega_0^2}\left\{-\frac{2\eta}{1-\eta^2}-\frac{2}{\eta}\cos\omega_0t\right.\\ &+\frac{1}{\eta(1+\eta)}\cos(1+\eta)\omega_0t\\ &\left.+\frac{1}{\eta(1-\eta)}\cos(1-\eta)\omega_0t\right\}\end{aligned}\right\}. \quad (5)$$

Figure 3, shows the projection on the transverse plane of the trajectory of a single electron for this case.

For $\kappa=\omega_0/\omega_c=1$, the maximum transverse deflection is obtained for both the linearly and circularly polarized fields. For a cavity of length λ_g, the electron gains energy during the first half-wave-length and loses it during the second half-wave-length.‡ If the transit time is chosen so that $1/\eta$ is an integer, all the rotational energy is lost during the second half wave-length. If $1/\eta$ is not an integer, there is a small amount of rotational energy left in the beam at exit. The deflection of the beam at $z=\lambda_g$ is of the order $1/\eta^2$ smaller than

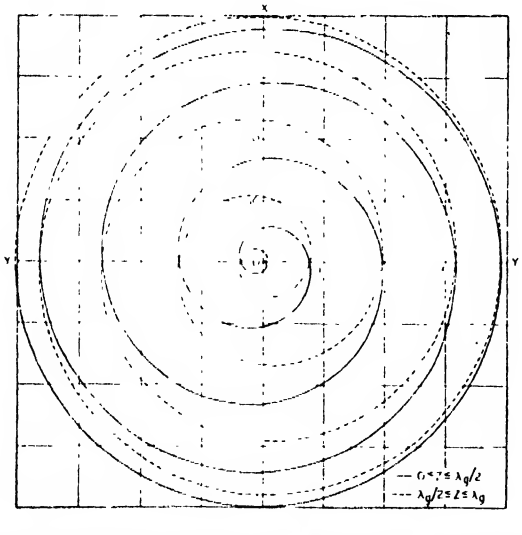


FIG. 3. Projection on the transverse plane of the trajectory of a single electron. TE_{11} cavity excited in a circularly polarized mode. $\kappa=1$, $\eta=0.1$.

‡ There are two unique cases, $\kappa=1/(1\pm\eta)$, for which there are equal increases in transverse deflection for each half wave-length traveled until relativistic effects become significant.

the deflection at $z=\lambda_g/2$. The deflection varies approximately as $1/\eta$. The transverse deflection for the circularly polarized case is approximately twice as great as the deflection for the linearly polarized case. For the same value of E_{0x}, it would require twice as much input power to drive the cavity in circular as in linear polarization.

For the $\kappa=1$ case, there are three sinusoidal terms in the expressions for $x(t)$ and $y(t)$ with frequencies $(1+\eta)\omega_0$, $(1-\eta)\omega_0$, and ω_0. The transit time, t_1, for a length $\lambda_g/2$ can be written as $\pi/\eta\omega_0$. Then, at time t_1 the arguments of these sinusoidal terms become

$$(1+\eta)\omega_0 t_1=\pi/\eta+\pi$$
$$(1-\eta)\omega_0 t_1=\pi/\eta-\pi$$
$$\omega_0 t_1=\pi/\eta.$$

The first two terms, $(1+\eta)\omega_0$ and $(1-\eta)\omega_0$, which describe the time dependence of the r-f driving E-field, differ by π-radians from their phase at $t=0$. Energy is gained by the electron in its transit from $z=0$ to $z=\lambda_g/2$ but is returned to the r-f field in the transit from $z=\lambda_g/2$ to $z=\lambda_g$.

By calculating the square root of the sum of x^2 and y^2 as given by Eq. (3), an expression is obtained for the radial deflection of the electron in terms of the entrance phase angle, ϕ, of the linearly polarized field. It is found that the terms independent of ϕ are of the order of $1/\eta^2$ larger than those containing ϕ. For values of $\eta\leq 0.1$, which are of the greatest practical interest, the radial deflection is effectively independent of ϕ. There is also a small spatial deviation of phase: while two electrons entering the cavity 90 degrees out-of-phase in time will still be 90 degrees out-of-phase in time at exit, they will not be 90 degrees out-of-phase in space. This bunching is negligible for values of $\eta\leq 0.1$.

For a circularly polarized mode, each electron has the same trajectory through the cavity except for the rotation of the coordinate system about the cylindrical axis.

In the solutions of the equations of motion for a single electron, the following approximations were made:

1. The E-field magnitude is constant for small deviations from the axis of the cavity.
2. The effect of the r-f B-field is negligible for velocities small with respect to the velocity of light and for small r-f power inputs.
3. The relativistic increase in electron mass is negligible.

For the choice of axes specified in Eqs. (2) for the linear mode, the expressions for the transverse E-field in a perfectly conducting TE_{11n_z} cavity are

$$\left.\begin{aligned} E_r&=-A_0(1/k_c r)J_1(k_c r)\cos\theta \\ E_\theta&=A_0 J_1'(k_c r)\sin\theta \end{aligned}\right\},$$

where J_1 = first-order Bessel function of first kind, J_1' = first derivative of J_1, $k_c=2\pi/\lambda_c$ where λ_c is the cut-off wave-length of the cavity, $A_0=2E_{0x}$ = amplitude constant.

Using the coordinate system of Fig. 1, the E-field can be written as

$$\left.\begin{aligned} E_x&=A_0\{[+(1/k_c r)J_1(k_c r) \\ &\qquad -\tfrac{1}{2}J_0(k_c r)]\cos 2\theta+\tfrac{1}{2}J_0(k_c r)\} \\ E_y&=A_0[-(1/k_c r)J_1(k_c r)+\tfrac{1}{2}J_0(k_c r)]\sin 2\theta. \end{aligned}\right\}$$

For small values of the argument $(k_c r)$,

$$(1/k_c r)J_1(k_c r)-\tfrac{1}{2}J_0(k_c r)\doteq 0$$

and these equations reduce to

$$E_x\doteq\tfrac{1}{2}A_0 J_0(k_c r)\doteq\tfrac{1}{2}A_0$$
$$E_y\doteq 0.$$

This approximation is valid to within 1.2 percent for values of the argument up to 0.3. Values of the argument up to 0.6 introduce errors of less than 5 percent. In a cavity designed to resonate at 9000 mc/sec., an argument of 0.5 corresponds to a deflection of one-quarter of the radius of the cavity, and, for $\omega_0=\omega_c$, corresponds to a tangential velocity of the electron roughly equal to the velocity of light.

The second approximation depends on the velocity of the electron and the r-f power input to the cavity. If the ratio of the transverse velocity of the electron is small with respect to c, the velocity of light, the force on the electron exerted by the E-field is large with respect to the force exerted by the B-field. If the power input to the cavity is large (above a kilowatt for an X band cavity), the magnitude of the r-f B-field approaches the magnitude of the steady B-field and therefore becomes appreciable.

The relativistic increase of mass is negligible if the ratio of the total velocity of the electron to the velocity of light is small with respect to one. For electrons of less than 5000 volts energy, relativistic effects introduce an error of less than 1 percent.

III. TRAJECTORY OF AN ELECTRON BEAM IN A TE_{11n_z} CAVITY IN A MAGNETIC FIELD

The theoretical treatment of the trajectory of a single electron in a cavity must be modified when a beam of electrons is considered. If the cavity is cylindrically symmetrical, two degenerate TE_{11n_z} modes can exist having mutually perpendicular orientations. The spiraling electron beam serves to couple energy between these two modes, and this coupling must be taken into account in the calculation of the beam trajectory.

The azimuthal position angle, θ, of the electrons in the cavity is related to the deflections by $\tan\theta=y/x$. For linear polarization with $\kappa=1$ and $\eta\ll 1$, the approximate expression for θ (after several r-f cycles) is

$$\theta\doteq\omega_c t+\phi+\pi/2.$$

In this expression each electron has its own local time which is taken to be zero when the electron enters the cavity.

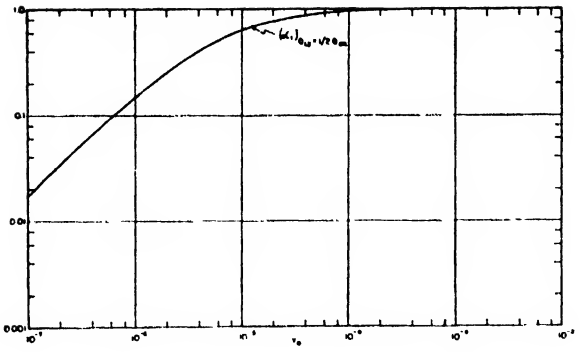


FIG. 4. $\alpha_1 = E_{2yx}/E_{1x}$ as a function of Y_0.
X band TE_{111} cavity with $\kappa = 1$.

If electron 1 enters at $\phi = 0$, its azimuthal angle (θ_1) is, after a time interval Δt_1, $\theta_1 = \omega_c \Delta t_1$. If electron 2 enters at $\phi = \phi_2$, its azimuthal angle (θ_2) is, after a time interval Δt_2, $\theta_2 = \omega_c \Delta t_2 + \phi_2$. Since $\phi_2 = \omega_c \Delta t_1 - \omega_c \Delta t_2$, $\theta_1 = \theta_2$. A projection on the transverse plane at any instant of a succession of electrons traversing the cavity would show a line of electrons each having the same azimuthal angle, but with greater radial deflections for the electrons which had been in the cavity for a longer interval of time. Accordingly, the beam of electrons has the form of a straight pencil at an angle to the axis of symmetry of the cavity rotating with an angular frequency ω_c.††

For circular polarization with $\kappa = 1$ and $\eta \ll 1$, the successive electrons will have the same azimuthal angle, as in the linear case. The beam has the same form of a straight pencil rotating around the axis of symmetry with an angular frequency ω_c.

For linear polarization, the spiraling electron beam drives a counter E-field in phase opposition to the original driving field. If the original r-f driving field is polarized in the plane containing the x axis, its amplitude E_{0x} is reduced by the counter E-field. This counter E-field is called E_{2xx} since it is directed along the x axis and is produced by the driving field in the plane of the x axis. With the reduction in the effective driving field, the losses in the cavity walls are decreased and the balance of the r-f input power appears in the rotational energy of the beam. The amplitude of the effective or reduced field is $|E_{1x}| = |E_{0x}| - |E_{2xx}|$.

The spiraling beam also excites the degenerate mode in the plane of the y axis. The amplitude of this cross-coupled field is called E_{2yx}. It is directed along the y axis, but is caused by the driving field in the plane of the x axis.

The differential equations of motion (1) for the electrons in the beam are unchanged, assuming the same approximations. For $\kappa = 1$ and $\eta^2 \ll 1$, the r-f field in the cavity as seen by the electron beam can be written as

$$\left.\begin{aligned} \mathcal{E}_{1x} &= E_{1x} \sin(2\pi z/\lambda_g) \cos\omega_0 t \\ \mathcal{E}_{2yx} &= -E_{2yx} \sin(2\pi z/\lambda_g) \sin\omega_0 t \\ \mathcal{E}_z &= 0 \end{aligned}\right\} \tag{6}$$

for the interval $0 \leq t \leq t_1$. The time dependence chosen for the E_{2yx} field is valid since the electron beam is a straight pencil beam rotating around the axis of symmetry. The phase angle of the E_{1x} field at entrance is omitted since $\eta^2 \ll 1$. The solution of the equations of motion is for $\kappa = 1$ (see Appendix A),

$$\left.\begin{aligned} x(t) = -\frac{eE_{1x}}{2m\omega_0^2}\Bigg[& \frac{4-2\eta^2}{\eta(4-\eta^2)} \sin\omega_0 t \\ & - \frac{\sin(1+\eta)\omega_0 t}{\eta(2+\eta)} - \frac{\sin(1-\eta)\omega_0 t}{\eta(2-\eta)} \\ & + \alpha_1 \Bigg\{ \frac{-4}{\eta(4-\eta^2)} \sin\omega_0 t + \frac{\sin(1+\eta)\omega_0 t}{\eta(2+\eta)(1+\eta)} \\ & \qquad + \frac{\sin(1-\eta)\omega_0 t}{\eta(2-\eta)(1-\eta)} \Bigg\} \Bigg] \\ y(t) = -\frac{eE_{1x}}{2m\omega_0^2}\Bigg[& \frac{-2\eta}{1-\eta^2} - \frac{4-2\eta^2}{\eta(4-\eta^2)} \cos\omega_0 t \\ & + \frac{\cos(1+\eta)\omega_0 t}{\eta(2+\eta)(1+\eta)} + \frac{\cos(1-\eta)\omega_0 t}{\eta(2-\eta)(1-\eta)} \\ & + \alpha_1 \Bigg\{ \frac{4}{\eta(4-\eta^2)} \cos\omega_0 t - \frac{\cos(1+\eta)\omega_0 t}{\eta(2+\eta)} \\ & \qquad - \frac{\cos(1-\eta)\omega_0 t}{\eta(2-\eta)} \Bigg\} \Bigg] \end{aligned}\right\} \tag{7}$$

This solution is written in terms of α_1 which is defined as the ratio of the E-field amplitudes, E_{2yx}/E_{1x}.

For a cavity excited in a circularly polarized mode, it is necessary to introduce E-field amplitudes for each plane of polarization corresponding to those defined for the linear case. The original driving fields in each plane have the amplitudes E_{0x} and E_{0y} when no energy is transferred to the beam. The reduced amplitudes are E_{1x} and E_{1y}.

There are two counter-fields in each plane: The first is the counter-field produced by the driving field in the same plane; the second is the counter-field cross-coupled from the driving field in the perpendicular plane. The first counter-field is analogous to the counter-e.m.f. in an electromechanical generator. The amplitude of this field is E_{2xx} in the x axis plane and E_{2yy} in the y axis plane. It acts in the same plane as the driving field by which it is produced.

The second type of counter-field, the cross-coupled field, acts in the plane perpendicular to the driving

†† A graphical analysis indicates that the $\eta \ll 1$ restriction is stronger then necessary and that a pencil beam is formed for $\eta^3 \ll 1$.

field by which it is produced. The field E_{2yx} acts in the y axis plane, but is caused by the driving field in the x axis plane. The field E_{2xy} acts in the x axis plane, but is caused by the driving field in the y axis plane.

Both types of counter-fields act in opposition to the phase of the driving fields if these original fields are chosen to add energy to the beam. Therefore

$$\begin{aligned} |E_{1x}| &= |E_{0x}| - |E_{2xx}| - |E_{2xy}| \\ |E_{1y}| &= |E_{0y}| - |E_{2yy}| - |E_{2yx}|. \end{aligned} \tag{8}$$

The differential equations of motion (1) are unchanged assuming the same approximations. For $\kappa=1$ and $\eta \ll 1$, the r-f field in the cavity as seen by the electron beam is

$$\left.\begin{aligned} \mathcal{E}_{1x} &= E_{1x} \sin(2\pi z/\lambda_g) \cos\omega_0 t \\ \mathcal{E}_{1y} &= E_{1x} \sin(2\pi z/\lambda_g) \sin\omega_0 t \\ \mathcal{E}_{1z} &= 0 \end{aligned}\right\} \tag{9}$$

for the interval $0 \leq t \leq t_1$. Except for the change in amplitude these expressions are identical with those for a single electron (4). Therefore, the solutions for a single electron (5) will be valid for the electron beam if E_{1x} is substituted for E_{0x}.

IV. TRAJECTORY OF A SPIRAL BEAM EXCITING A TE_{11n_z} CAVITY IN A MAGNETIC FIELD

If a spiraling beam of electrons is projected into a TE_{11n_z} cavity in the presence of a steady axial magnetic field, the cavity is excited, and energy is extracted from the electron beam. It is assumed in this discussion that a steady state condition has been reached. It is also assumed that the spiraling beam is a straight pencil of electrons rotating about the axis of symmetry at an angular frequency ω_c. The phase of the E-field in the cavity is determined by the phase of the electrons upon entrance ($t=0$). If the phase of the entering beam is chosen so that $x_0=0$, $\dot{y}_0=0$, $\dot{x}_0$ has its maximum negative value, and y_0 has its maximum positive value, then the r-f field as seen by the electron beam for $\kappa=1$ and $\eta^2 \ll 1$ is

$$\left.\begin{aligned} \mathcal{E}_{2x} &= -E_{2x} \sin(2\pi z/\lambda_g) \cos\omega_0 t \\ \mathcal{E}_{2y} &= -E_{2y} \sin(2\pi z/\lambda_g) \sin\omega_0 t \\ \mathcal{E}_{2z} &= 0 \end{aligned}\right\} \tag{10}$$

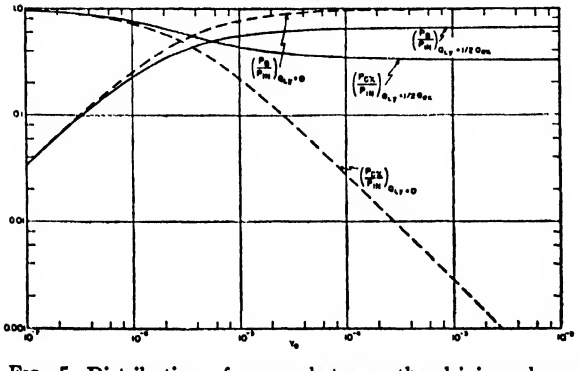


FIG. 5. Distribution of power between the driving plane of polarization and the electron beam. X band TE_{111} cavity with $\kappa=1$.

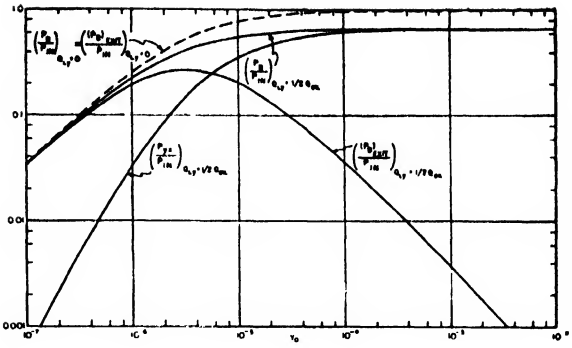


FIG. 6. Distribution of power between the electron beam and the driven plane of polarization. X band TE_{111} cavity with $\kappa=1$.

for the interval $0 \leq t \leq t_1$. The trajectory of the beam can be obtained from the solution of the differential equations of motion (1). The solution is for $\kappa=1$ (see Appendix B),

$$\begin{aligned} x(t) = {} & \frac{\dot{x}_0}{\omega_0} \sin\omega_0 t + \frac{eE_{2x}}{2m\omega_0^2} \left\{ \frac{4-2\eta^2}{\eta(4-\eta^2)} \sin\omega_0 t \right. \\ & \left. - \frac{\sin(1+\eta)\omega_0 t}{\eta(2+\eta)} - \frac{\sin(1-\eta)\omega_0 t}{\eta(2-\eta)} \right\} \\ & + \frac{eE_{2y}}{2m\omega_0^2} \left\{ \frac{4}{\eta(4-\eta^2)} \sin\omega_0 t \right. \\ & \left. - \frac{\sin(1+\eta)\omega_0 t}{\eta(1+\eta)(2+\eta)} - \frac{\sin(1-\eta)\omega_0 t}{\eta(1-\eta)(2-\eta)} \right\} \\ y(t) = {} & y_0 + \frac{\dot{x}_0}{\omega_0} - \frac{\dot{x}_0}{\omega_0} \cos\omega_0 t \\ & + \frac{eE_{2x}}{2m\omega_0^2} \left\{ -\frac{2\eta}{1-\eta^2} - \frac{4-2\eta^2}{\eta(4-\eta^2)} \cos\omega_0 t \right. \\ & \left. + \frac{\cos(1+\eta)\omega_0 t}{\eta(1+\eta)(2+\eta)} + \frac{\cos(1-\eta)\omega_0 t}{\eta(1-\eta)(2-\eta)} \right\} \\ & + \frac{eE_{2y}}{2m\omega_0^2} \left\{ \frac{-4}{\eta(4-\eta^2)} \cos\omega_0 t \right. \\ & \left. + \frac{\cos(1+\eta)\omega_0 t}{\eta(2+\eta)} + \frac{\cos(1-\eta)\omega_0 t}{\eta(2-\eta)} \right\} \end{aligned} \tag{11}$$

V. COUPLING BETWEEN OSCILLATIONS IN MUTUALLY PERPENDICULAR PLANES OF POLARIZATION IN A CAVITY DRIVEN IN LINEAR POLARIZATION

The power input (P_{in}) to the cavity excited in linear polarization in the x axis plane is partially dissipated in the walls of the cavity (P_{cx}) and partially transferred to the electron beam (P_B). P_B is determined from

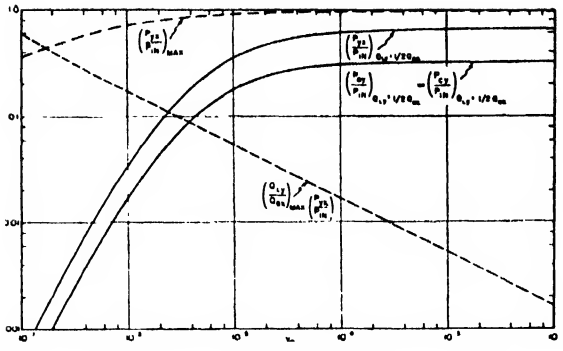


FIG. 7. Distribution of power in the plane of polarization driven by the electron beam. X band TE_{111} cavity with $\kappa=1$.

the beam current and the driving field along the trajectory of the beam,

$$P_B=\int \text{beam current} \times \text{driving field}.$$

The trajectory of the beam of electrons in a linearly excited cavity is given in Section III in terms of E_{1x} and $\alpha_1=E_{2yx}/E_{1x}$.

P_{cx} is obtained with the usual expression for Q,

$$Q=\omega_0\frac{\text{Energy stored}}{\text{Average power loss}},$$

in terms of P_{in}, E_{1x}, and the properties of the cavity. From these two expressions, P_B and P_{cx} can be written in terms of α_1 and the physical constants of the cavity.

To find α_1, the power coupled from the electron beam to the second plane of polarization, P_{yx}, must be evaluated. P_{yx} is found by taking the line integral of the product of the beam current and the driven E-field, E_{2yx}, along the trajectory of the beam. With the expression for the Q in the driven plane, another relation between E_{2yx} and P_{yx} is available. These two expressions can be solved simultaneously to give α_1 in terms of the beam admittance, $Y_0=I_0/V_0$, and cavity constants. With α_1 determined, P_B, P_{cx}, P_{yx}, and the rotational power left in the beam at exit can be written in terms of the r-f power input, P_{in}. If there is an iris coupled to the second plane of polarization, the power transmitted through the iris (P_{0y}) and the power dissipated in the cavity walls in the y axis plane (P_{cy}) can also be determined.

In the following treatment, α_1 is obtained first. The general power expressions are specialized to the case of a cavity of length $\lambda_g/2$ with $\kappa=1$ and $\eta^2\ll1$. The resulting expressions are plotted for a 9000-mc/sec. (X band) TE_{111} cavity.

The power coupled to the plane containing the y axis is given by

$$P_{yx}=Ne\int \boldsymbol{E}_2\cdot\mathbf{ds}, \tag{12}$$

where $N=$number of electrons entering per second,

$\boldsymbol{E}_2=-y'E_{2yx}\sin2\pi z/\lambda_g\sin\omega_0 t$,

$\mathbf{ds}=\boldsymbol{x}\dot{x}dt+\boldsymbol{y}\dot{y}dt+\boldsymbol{z}\dot{z}dt=$path element of the trajectory.

Since both ξ_2 and $\mathbf{s}$ can be written as functions of t only, using $z=v_0t$, the limits of integration can be taken as $t=0$ and $t=t_1$, where t_1 is the transit time of an electron through the cavity.

For the case of a cavity of length $\lambda_g/2$ with $\kappa=1$, expression (12) becomes with (7) since $\eta^2\ll1$,

$$P_{yx}=\frac{I_0eE_{2xy}{}^2}{2m\omega_0{}^2\eta^2}\frac{(1-\alpha_1)}{\alpha_1}=\frac{Y_0\lambda_g{}^2E_{2yx}{}^2}{16\pi^2}\frac{(1-\alpha_1)}{\alpha_1}, \tag{13}$$

where $I_0=Ne=$beam current, $Y_0=I_0/V_0=$beam admittance, and $V_0=mv_0{}^2/2e=$beam accelerating voltage. With the usual definition of the cavity Q,

$$\left.\begin{aligned}P_{yx}&=\omega_0\left(\tfrac{1}{2}\epsilon_0\int E_2{}^2d\tau/Q_{Ly}\right)\\&=\omega_0\epsilon_0E_{2yx}{}^2\tau_0/2Q_{Ly}\end{aligned}\right\}, \tag{14}$$

where $Q_{Ly}=$the y plane Q loaded by the coupling iris, $\epsilon_0=$dielectric constant of free space, $d\tau=$element of volume in the cavity, $\tau_0=$volume constant (see Appendix C). From expressions (13) and (14), the desired result is obtained:

$$\alpha_1=\frac{1}{1+(8\pi^2\omega_0\epsilon_0\tau_0)/(Y_0Q_{Ly}\lambda_g{}^2)}. \tag{15}$$

To relate P_{yx} to P_{in}, it is first necessary to evaluate the power coupled into the electron beam, P_B.

$$P_B=-Ne\int \boldsymbol{E}_1\cdot\mathbf{ds}, \tag{16}$$

where $E'=\boldsymbol{x}E_{1x}\sin(2\pi z/\lambda_g)\cos\omega_0t$. For the case of a cavity of length $\lambda_g/2$, with $\kappa=1$ since $\eta^2\ll1$,

$$P_B=\frac{I_0eE_{1x}{}^2}{2m\omega_0{}^2}\frac{1}{\eta^2}(1-\alpha_1)=\frac{Y_0\lambda_g{}^2E_{1x}{}^2}{16\pi^2}(1-\alpha_1). \tag{17}$$

The power driving the cavity, P_{in}, is partially dissipated in the walls of the cavity and the rest is converted into rotational energy of the electrons. The power dissipated in the walls of the cavity in the x axis plane, P_{cx}, can be written as

$$P_{cx}=\omega_0\epsilon_0E_{1x}{}^2\tau_0/2Q_{0x}, \tag{18}$$

where $Q_{0x}=$unloaded Q of the mode in the x axis plane. Defining k_1 as the ratio P_B/P_{cx},

$$k_1=P_B/P_{cx}=Y_0Q_{0x}\lambda_g{}^2(1-\alpha_1)/8\pi^2\omega_0\epsilon_0\tau_0 \tag{19}$$

for the case of $\kappa=1$ and a cavity length of $\lambda_g/2$. Sub-

stituting the expression for α_1, (15) results in

$$k_1=(Q_{0x}/Q_{Ly})\alpha_1. \tag{20}$$

Since $P_{in}=P_{cx}+P_B$,

$$\left.\begin{aligned} P_B&=\frac{k_1}{1+k_1}P_{in}=\frac{1}{1+(Q_{Ly})/(Q_{0x}\alpha_1)}P_{in}\\ \text{and}\qquad\qquad&\\ P_{cx}&=\frac{1}{1+k_1}P_{in}=\frac{1}{1+(Q_{0x}/Q_{Ly})\alpha_1}P_{in}\end{aligned}\right\}. \tag{21}$$

By defining two more coupling coefficients

$$k_2=P_{yx}/P_B \quad \text{and} \quad k_3=P_{0y}/P_{yx},$$

where P_{0y} is the power coupled out through an iris in the y axis plane, it is possible to write P_{0y} in terms of P_{in}.

The coefficient k_3 depends on the properties of the coupling iris and can be written as

$$k_3=1-Q_{Ly}/Q_{0y}, \tag{22}$$

where Q_{0y} is the unloaded Q in the y axis plane. From expressions (13) and (17)

$$k_2=\alpha_1$$

and

$$\frac{P_{yx}}{P_{in}}=\frac{\alpha_1}{1+(Q_{Ly})/(Q_{0x}\alpha_1)}. \tag{23}$$

Then, by the definitions of the coupling coefficients

$$\frac{P_{0y}}{P_{in}}=\frac{k_1k_2k_3}{1+k_1}=\frac{\alpha_1}{1+(Q_{Ly})/(Q_{0x}\alpha_1)}\left(1-\frac{Q_{Ly}}{Q_{0y}}\right). \tag{24}$$

Thus, the ratio P_{0y}/P_{in} is specified in terms of the beam admittance and the physical characteristics of the cavity.

The rotational power remaining in the spiral beam at the cavity exit, $(P_B)_{exit}$, can be obtained by subtracting P_{yx} from P_B. It is also possible to obtain $(P_B)_{exit}$ from kinetic energy considerations:

$$\left.\begin{aligned}(P_B)_{exit}&=Nm/2[(\dot{x})^2_{t=t_1}+(\dot{y})^2_{t=t_1}]\\ &=\frac{I_0eE_{1x}^2}{2m\omega_0^2\eta^2}(1-\alpha_1)^2=\frac{1-\alpha_1}{1+(Q_{Ly})/(Q_{0x}\alpha_1)}P_{in}\end{aligned}\right\} \tag{25}$$

for a cavity of length $\lambda_0/2$ with $\kappa=1$ and $\eta^2\ll 1$.

The amplitudes of the E-fields defined in Section III can now be readily evaluated in terms of P_{in} for a cavity of length $\lambda_0/2$ with $\kappa=1$. By definition of Q_{0x},

$$E_{0x}=(2Q_{0x}P_{in}/\omega_0\epsilon_0\tau_0)^{\frac{1}{2}}. \tag{26}$$

From (18) and (21),

$$E_{1x}=\left[\frac{2Q_{0x}}{\omega_0\epsilon_0\tau_0}\frac{P_{in}}{1+(Q_{0x}/Q_{Ly})\alpha_1}\right]^{\frac{1}{2}}. \tag{27}$$

From (27) and the definition of α_1,

$$E_{2yx}=\left[\frac{2Q_{0x}}{\omega_0\epsilon_0\tau_0}\frac{\alpha_1^2P_{in}}{1+(Q_{0x})/(Q_{Ly})\alpha_1}\right]^{\frac{1}{2}}. \tag{28}$$

The "counter" E-field, E_{2xx} defined in Section III, is obtained from (26) and (27),

$$E_{2xx}=\left\{\frac{2Q_{0x}P_{in}}{\omega_0\epsilon_0\tau_0}\left[1-\frac{1}{1+(Q_{0x}/Q_{Ly})\alpha_1}\right]\right\}^{\frac{1}{2}}. \tag{29}$$

The actual field present in the cavity is elliptically polarized and rotates in the sense opposite to that of the electrons.

In order to plot some of the foregoing expressions as functions of Y_0, the following values are assumed for an X band TE_{111} cavity with two coupling holes:

$$\begin{aligned}\omega_0&=2\pi\times 9\times 10^9 \text{ c.p.s.},\\ \lambda_g&=4\times 10^{-2} \text{ meter},\\ Q_{0x}&=Q_{0y}=2\times 10^4,\\ Q_{Ly}&=10^4,\\ \tau_0&=2.28\times 10^{-6} \text{ meter}^3\end{aligned}$$

(see Appendix C). All plots are for $\eta^2\ll 1$ and $\kappa=1$, which corresponds to a B-field of 0.32 weber/meter^2 (3200 gauss). A practical range of values for the beam admittance, Y_0, is 10^{-2} to 10^{-7}: $Y_0=10^{-2}$ mhos corresponds to currents of the order of 10 or 100 ma, and voltages of the order of 1 to 10 volts; $Y_0=10^{-7}$ mhos corresponds to currents of the order of 10μa, and voltages of the order of 100 volts.

Figure 4 shows a plot of $\alpha_1=E_{2yx}/E_{1x}$ as a function of Y_0, Eq. (15).

Figure 5 shows a plot of the distribution of power ($P_{in}=P_B+P_{cx}$) between the driving (x axis) plane of polarization and the electron beam as a function of Y_0.

If the mode in the y axis plane is suppressed ($Q_{Ly}=0$), the approximate expressions (21) become indeterminate and it is necessary to let α_1 vanish in the expression for P_B (17). Then

$$P_B=Y_0\lambda_g^2E_{1x}^2/16\pi^2. \tag{30}$$

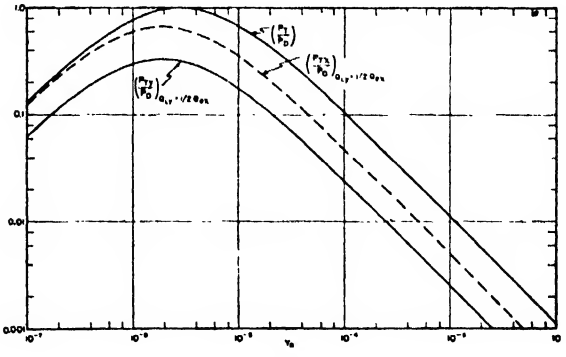


FIG. 8. Power transfer from a spiraling beam of electrons to an X band TE_{111} cavity. $\kappa=1$ and $Q_{Lx}\neq Q_{Ly}$.

Since P_{cx} remains the same,

$$\left.\begin{aligned}\frac{P_B}{P_{\text{in}}}&=\frac{1}{1+(8\pi^2\omega_0\epsilon_0\tau_0)/(Y_0\lambda_g^2Q_{0x})}\\ \text{and}\qquad\\ \frac{P_{cx}}{P_{\text{in}}}&=\frac{1}{1+(Y_0\lambda_g^2Q_{0x})/(8\pi^2\omega_0\epsilon_0\tau_0)}\end{aligned}\right\}. \tag{31}$$

These curves are also shown on Fig. 5.

Figure 6 shows the distribution of the power coupled into the beam ($P_B=P_{yx}+P_B$) as a function of Y_0. If the mode in the y axis plane is suppressed, ($Q_{Ly}=0$), $P_{yx}=0$, and $(P_B)_{\text{exit}}=P_B$.

Figure 7 shows the distribution of power in the driven (y axis) plane of polarization ($P_{yx}=P_{cy}+P_{0y}$) as a function of Y_0. P_{cy} is the power dissipated in the walls of the cavity by the oscillation in the y axis plane and can be written

$$\left.\begin{aligned}P_{cy}&=\frac{\omega_0\epsilon_0\tau_0E_{2yx}^2}{2Q_{0y}}\\ &=\frac{\alpha_1}{1+((Q_{Ly})/(Q_{0x}\alpha_1)}\frac{Q_{Ly}}{Q_{0y}}P_{\text{in}}\end{aligned}\right\} \tag{32}$$

by making use of (28). Since critical coupling ($Q_{0y}=2Q_{Ly}$) was chosen, $P_{0y}=P_{cy}$. It is of interest to determine the dependence of P_{yx} as a function of Q_{Ly}. Using expression (15) for α_1 and differentiating the power ratio P_{yx}/P_{in} (23), with respect to Q_{Ly}/Q_{0x}, the following condition results for maximum P_{yx}/P_{in},

$$\frac{Q_{Ly}}{Q_{0x}}=\left[\frac{8\pi^2\omega_0\epsilon_0\tau_0}{\lambda_g^2Y_0Q_{0x}}+\left(\frac{8\pi^2\omega_0\epsilon_0\tau_0}{\lambda_g^2Y_0Q_{0x}}\right)^2\right]^{\frac{1}{2}}. \tag{33}$$

Substituting this expression in (23), the optimum power ratio P_{yx}/P_{in} is obtained. This optimum power ratio and the ratio Q_{Ly}/Q_{0x} are also plotted in Fig. 7.

The value of Q_{Ly} can be decreased by increasing the size of the coupling hole in the y axis plane. Since it is not possible to enlarge the hole very much beyond critical coupling without materially decreasing the values of Q_{0y} and Q_{0x}, the maximization condition (33) has a limited validity.

VI. COUPLING BETWEEN OSCILLATIONS IN MUTUALLY PERPENDICULAR PLANES OF POLARIZATION IN A CAVITY DRIVEN IN CIRCULAR POLARIZATION

Since all the power in the beam appears at exit for the case of circular polarization, $P_B=(P_B)_{\text{exit}}$. From kinetic energy considerations,

$$P_B=Nm/2[(\dot{x})^2{}_{t=t_1}+(\dot{y})^2{}_{t=t_1}],$$

where $\dot{x}$ and $\dot{y}$ are found by differentiating expressions (5) with E_{1x} substituted for E_{0x}. For a cavity of length $\lambda_g/2$ with $\kappa=1$ for $\eta\ll1$,

$$P_B=(Y_0E_{1x}^2\lambda_g^2)/4\pi^2. \tag{34}$$

For a cylindrically symmetrical cavity, one-half of the power in the beam is coupled from the driving field in each of the two planes of polarization. Thus it is possible to define coefficients, similar to k_1 (19),

$$k_{1x}=P_{Bx}/P_{cx}$$

and

$$k_{1y}=P_{By}/P_{cy},$$

where $P_{Bx}=P_{By}=\frac{1}{2}P_B$. Then $k_{1x}=k_{1y}$. Using (18) for P_{cx}

$$k_{1x}=\frac{Y_0\lambda_g^2Q_{0x}}{4\pi^2\omega_0\epsilon_0\tau_0} \tag{35}$$

is obtained. By analogy with the expressions (21),

$$\left.\begin{aligned}P_{Bx}&=[k_{1x}/(1+k_{1x})](P_{\text{in}})_x\\ P_{cx}&=[1/(1+k_{1x})](P_{\text{in}})_x\end{aligned}\right\}. \tag{36}$$

Similar expressions hold for the y axis plane.

From (36) and (18), E_{1x} can be written explicitly as

$$E_{1x}=\left[\frac{2Q_{0x}}{\omega_0\epsilon_0\tau_0}\frac{(P_{\text{in}})_x}{1+k_{1x}}\right]^{\frac{1}{2}} \tag{37}$$

and by definition

$$E_{0x}=\left[\frac{2Q_{0x}}{\omega_0\epsilon_0\tau_0}(P_{\text{in}})_x\right]^{\frac{1}{2}}.$$

From expression (8),

$$E_{2xx}+E_{2xy}=E_{0x}-E_{1x}=\left[\frac{2Q_{0x}(P_{\text{in}})_x}{\omega_0\epsilon_0\tau_0}\left(\frac{k_{1x}}{1+k_{1x}}\right)\right]^{\frac{1}{2}}. \tag{38}$$

Similar expressions hold for the y axis plane.

VII. EXCITATION OF A TE_{11n_z} CAVITY BY A SPIRALING BEAM OF ELECTRONS

Since the transient case is of little practical interest, it is assumed that steady-state conditions have been reached.

In Section IV, the expressions for the trajectory of the electron beam were determined for the case of a cavity excited by a spiraling beam. The power transferred to the cavity is found by taking the line integral of the product of the beam current and the E-field along the trajectory of the beam through the cavity. If the Q's of the two modes are not the same, it is convenient to evaluate separately the line integrals of the x and y components of the r-f field. The general expressions are specialized to the case of a cavity of length $\lambda_g/2$ with $\kappa=1$ and $\eta^2\ll1$.

The power transferred to the cavity, P_T, is given by the integral

$$P_T = +Ne\int \boldsymbol{E}_2\cdot\mathbf{ds}, \tag{39}$$

where

$$\boldsymbol{E}_2 = -\boldsymbol{x}E_{2x}\sin(2\pi z/\lambda_g)\cos\omega_0 t - \boldsymbol{y}E_{2y}\sin(2\pi z/\lambda_g)\sin\omega_0 t$$

and

$$\mathbf{ds} = \boldsymbol{x}\dot{x}dt + \boldsymbol{y}\dot{y}dt.$$

Considering first the case where $Q_{Lx}\neq Q_{Ly}$, it is convenient to write P_T as the sum $(P_{Tx}+P_{Ty})$ of the powers transferred to the two modes. Then

$$\left.\begin{aligned} P_{Tx} &= -Ne\int \boldsymbol{x}E_{2x}\cdot\mathbf{ds} \\ \text{and} \\ P_{Ty} &= -Ne\int \boldsymbol{y}E_{2y}\cdot\mathbf{ds} \end{aligned}\right\}. \tag{40}$$

The evaluation of the first of these integrals for a cavity of length $\lambda_g/2$ with $\kappa=1$ results in since $\eta^2\ll 1$,

$$P_{Tx} = -\frac{NeE_{2x}}{\omega_0\eta}\dot{x}_0 - \frac{Ne^2E_{2x}^2}{2m\omega_0^2\eta^2} - \frac{Ne^2E_{2x}E_{2y}}{2m\omega_0^2\eta^2}. \tag{41}$$

With the usual expression for the cavity Q,

$$P_{Tx} = \omega_0\epsilon_0\tau_0 E_{2x}^2/2Q_{Lx}. \tag{42}$$

These expressions determine E_{2x},

$$E_{2x} = \frac{-2I_0\dot{x}_0/\omega_0\eta}{(\omega_0\epsilon_0\tau_0/Q_{Lx}) + (Y_0\lambda_g^2/8\pi^2)(1+Q_{Ly}/Q_{Lx})} \tag{43}$$

where $\dot{x}_0$ is negative. Using (42)

$$P_{Tx} = \frac{\omega_0\epsilon_0\tau_0}{2Q_{Lx}} \times\left[\frac{-2I_0\dot{x}_0/\omega_0\eta}{(\omega_0\epsilon_0\tau_0/Q_{Lx}) + (Y_0\lambda_g^2/8\pi^2)(1+Q_{Ly}/Q_{Lx})}\right]^2. \tag{44}$$

Defining the driving power as

$$P_D = (Nm/2)\dot{x}_0^2, \tag{45}$$

it is possible to write (44) as the power ratio,

$$\frac{P_{Tx}}{P_D} = \frac{Q_{Lx}Y_0\lambda_g^2}{2\pi^2\omega_0\epsilon_0\tau_0} \times\left[\frac{1}{1+[Y_0\lambda_g^2(Q_{Lx}+Q_{Ly})]/[8\pi^2\omega_0\epsilon_0\tau_0]}\right]^2. \tag{46}$$

Similar expressions hold for the y axis plane.

If the cavity has one iris which couples to the y axis plane, Q_{Lx} will be greater than Q_{Ly}, and more power is coupled to the unloaded plane of polarization.

For the special case where $Q_{Lx}=Q_{Ly}$, $E_{2x}=E_{2y}$ and the total power transferred is $2P_{Tx}$,

$$\frac{P_T}{P_D} = \frac{Q_{Lx}Y_0\lambda_g^2}{\pi^2\omega_0\epsilon_0\tau_0}\left[\frac{1}{1+[Q_{Lx}Y_0\lambda_g^2]/[4\pi^2\omega_0\epsilon_0\tau_0]}\right]^2. \tag{47}$$

Figure 8 shows the power ratio P_T/P_D (47) as a function of Y_0 for a cylindrically symmetrical X band cavity having the same physical constants as used in Section V with Q_{Lx} taken equal to Q_{Ly}. Above the critical value of Y_0 for which P_T/P_D equals unity, the power transferred to the cavity drops off rapidly. The physical explanation is that for higher values of Y_0, the electrons lose all their rotational energy to the r-f field before they complete their transit through the cavity, and start to regain rotational energy from the field.

Figure 8 also shows P_{Tx}/P_D (46) and P_{Ty}/P_D (46) plotted as functions of Y_0 for the case where there is a single iris critically coupled to the y axis plane. Then

$$Q_{Lx} = Q_{0x} = 2Q_{Ly}.$$

The total power transferred to the beam $(P_{Tx}+P_{Ty})$ is

$$\frac{P_T}{P_D} = (Q_{Lx}+Q_{Ly})\frac{Y_0\lambda_g^2}{2\pi^2\omega_0\epsilon_0\tau_0} \times\left[\frac{1}{1+[Y_0\lambda_g^2(Q_{Lx}+Q_{Ly})]/[8\pi^2\omega_0\epsilon_0\tau_0]}\right]^2.$$

Maximizing this ratio with respect to Y_0 determines the critical value of Y_0,

$$Y_0 = \frac{8\pi^2\epsilon_0\omega_0\tau_0}{\lambda_g^2(Q_{Lx}+Q_{Ly})}.$$

VIII. CONCLUSIONS

Several possible applications of the interaction between the resonant cavity oscillations and the spiral beam of electrons are evident. First, this interaction could be used as the basic energy interchange mechanism in a microwave generator. Second, the coupling between oscillations in the two planes of polarization in a single cavity or the coupling between two cavities by an electron beam could serve as a method of amplitude modulating a microwave signal. Modulation could be effected by varying either the beam current or voltage. A tube called the Electron Coupler, which uses a two-cavity system for amplitude modulation, has

recently been described in the literature.[7] Third, the resonant cavity driven by an external oscillator could be used as a microwave circular sweep generator. This sweep would be particularly useful for measuring extremely short pulses. For example, a spiral beam formed in the X band cavity described would make one complete revolution in about 10^{-10} second and could therefore be used to measure time intervals up to the order of 10^{-11} second.

Experimental X band tubes have been constructed in this Laboratory. The experimental results obtained will be published in a later paper.

APPENDIX A

Solution of the Equations of Motion in a TE_{11n_z} Cavity

The equations of motion (1) can be written using the Laplace transform notation[8] for $0 \leq t \leq t_1$,

$$\left.\begin{aligned} s^2X(s)+\omega_c sY(s) &= -(e/m)\mathfrak{L}(\mathcal{E}_x) \\ s^2Y(s)-\omega_c sX(s) &= -(e/m)\mathfrak{L}(\mathcal{E}_y) \end{aligned}\right\},$$

where s=variable in the complex plane, $\mathfrak{L}$=Laplace transform of, $X(s)$=Laplace transform of $x(t)$, $Y(s)$=Laplace transform of $y(t)$.

The initial conditions assumed are that $x=0$, $\dot{x}=0$, $y=0$, and $\dot{y}=0$ at $t=0$. Then

$$\left.\begin{aligned} X(s) = &-\frac{eE_{0x}}{2m}\frac{\omega_c}{s(s^2+\omega_c^2)} \\ &\times\mathfrak{L}\{\cos[(1+\eta)\omega_0 t+\phi]-\cos[(1-\eta)\omega_0 t+\phi]\} \\ &-\frac{eE_{0y}}{2m}\frac{1}{s^2+\omega_c^2} \\ &\times\mathfrak{L}\{\sin[(1+\eta)\omega_0 t+\phi]-\sin[(1-\eta)\omega_0 t+\phi]\} \end{aligned}\right\}$$

and

$$\left.\begin{aligned} Y(s) = &-\frac{eE_{0x}}{2m}\frac{\omega_c}{s(s^2+\omega_c^2)} \\ &\times\mathfrak{L}\{\sin[(1+\eta)\omega_0 t+\phi]-\sin[(1-\eta)\omega_0 t+\phi]\} \\ &+\frac{eE_{0y}}{2m}\frac{1}{s^2+\omega_c^2} \\ &\times\mathfrak{L}\{\cos[(1+\eta)\omega_0 t+\phi]-\cos[(1-\eta)\omega_0 t+\phi]\} \end{aligned}\right\}.$$

For the linearly polarized mode, $E_{0y}=0$. Then, for $\omega_0 \neq (1\pm\eta)\omega_0$,

$$\left.\begin{aligned} x(t) = &-\frac{eE_{0x}}{2m}\mathfrak{L}^{-1}\left[\frac{(1+\eta)\omega_0\cos\phi+s\sin\phi}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.-\frac{(1-\eta)\omega_0\cos\phi+s\sin\phi}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right] \\ y(t) = &-\frac{eE_{0x}}{2m}\omega_c\mathfrak{L}^{-1}\left[\frac{(1+\eta)\omega_0\cos\phi+s\sin\phi}{s(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.-\frac{(1-\eta)\omega_0\cos\phi+s\sin\phi}{s(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right] \end{aligned}\right\},$$

where $\mathfrak{L}^{-1}$ means the inverse Laplace transform. By taking the inverse and letting $\omega_0=\omega_c$, i.e., $\kappa=1$, the Eqs. (3) are obtained.

For the circularly polarized driving field, $E_{0y}\neq 0$ and $\phi=0$, then

$$\left.\begin{aligned} x(t) = &-\frac{eE_{0x}}{2m}\mathfrak{L}^{-1}\left[\frac{(1+\eta)\omega_0}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.-\frac{(1-\eta)\omega_0}{(s^2+\omega_c^2)(s^2-(1-\eta)^2\omega_0^2)}\right] \\ &-\frac{eE_{0y}}{2m}\mathfrak{L}^{-1}\left[\frac{\omega_c}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.-\frac{\omega_c}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right] \\ y(t) = &-\frac{eE_{0x}}{2m}\mathfrak{L}^{-1}\left[\frac{\omega_c(1+\eta)\omega_0}{s(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.-\frac{\omega_c(1-\eta)\omega_0}{s(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right] \\ &-\frac{eE_{0y}}{2m}\mathfrak{L}^{-1}\left[\frac{s}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}\right. \\ &\left.+\frac{s}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right] \end{aligned}\right\}.$$

By taking the inverse of these expressions, and letting $\omega_0=\omega_c$, i.e., $\kappa=1$, and $E_{0x}=E_{0y}$, Eqs. (5) are obtained. Equations (7) are obtained when E_{1x} is substituted for E_{0x} and $-E_{2yx}$ is substituted for E_{0y}.

APPENDIX B

Solution of the Equations of Motion in a TE_{11n_z} Cavity Driven by a Spiraling Electron Beam

The equations of motion (1) can be written in Laplace transform notation[8] for $0 \leq t \leq t_1$,

$$\left.\begin{aligned} s^2X(s)-\dot{x}_0+\omega_c sY(s)-\omega_c y_0 &= -(e/m)\mathfrak{L}(\mathcal{E}_x) \\ s^2Y(s)-y_0 s-\omega_c sX(s) &= -(e/m)\mathfrak{L}(\mathcal{E}_y) \end{aligned}\right\},$$

[7] C. L. Cuccia, "The electron coupler—A developmental tube for amplitude modulation and power control at ultra-high frequencies," RCA Rev. Vol. X, 2, 270-303 (1949).

[8] M. F. Gardner and J. L. Barnes, *Transients in Linear Systems* (John Wiley and Sons, Inc., New York, 1942).

where the initial conditions at $t=0$ are: $x=0$, $y=y_0$, $\dot{x}=\dot{x}_0$, $\dot{y}=0$. Then,

$$x(t)=\mathfrak{L}^{-1}\left\{\frac{\dot{x}_0}{s^2+\omega_c^2}+\frac{eE_{2x}}{2m}\left[\frac{(1+\eta)\omega_0}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}-\frac{(1-\eta)\omega_0}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right]\right.$$
$$\left.+\frac{eE_{2y}}{2m}\left[\frac{\omega_c}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}-\frac{\omega_c}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right]\right\}$$

$$y(t)=\mathfrak{L}^{-1}\left\{\frac{y_0 s}{s^2+\omega_c^2}+\frac{\omega_c^2 y_0+\omega_c\dot{x}_0}{s(s^2+\omega_c^2)}-\frac{eE_{2x}}{2m}\left[\frac{\omega_c(1+\eta)\omega_0}{s(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}-\frac{\omega_c(1-\eta)\omega_0}{s(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right]\right.$$
$$\left.+\frac{eE_{2y}}{2m}\left[\frac{s}{(s^2+\omega_c^2)(s^2+(1+\eta)^2\omega_0^2)}-\frac{s}{(s^2+\omega_c^2)(s^2+(1-\eta)^2\omega_0^2)}\right]\right\}.$$

By taking the inverse and letting $\omega_0=\omega_c$, i.e., $\kappa=1$, Eqs. (11) are obtained.

APPENDIX C

Evaluation of $\frac{1}{2}\epsilon_0\int E^2 d\tau$ in TE_{11n_z} Cavity

The E-field in a TE_{11n_z} cavity is given by[9]

$$\left.\begin{aligned}E_r&=-A_0\frac{1}{k_c r}J_1(k_c r)\sin\theta\sin\frac{2\pi n_z z}{\lambda_g}\\ E_\theta&=A_0 J_1'(k_c r)\cos\theta\sin\frac{2\pi n_z z}{\lambda_g}\end{aligned}\right\},\qquad\text{(C-1)}$$

where n_z is an integer. Then

$$\frac{1}{2}\epsilon_0\int E^2 d\tau=\frac{\epsilon_0 A_0^2}{2}\int_0^{(n_z\lambda_g/2)}\sin^2\frac{2\pi n_z z}{\lambda_g}dz\times\int_0^a\int_0^{2\pi}\left[\frac{1}{k_c^2 r^2}J_1^2(k_c r)\sin^2\theta+J_1'^2(k_c r)\cos^2\theta\right]r\,dr\,d\theta$$
$$=\frac{\epsilon_0 A_0^2}{2}\frac{n\lambda_g\pi}{4k_c^2}\left[\int_0^{k_c a}\frac{1}{\rho}J_1^2(\rho)d\rho+\int_0^a\rho J_1'^2(\rho)d\rho\right],\qquad\text{(C-2)}$$

[9] S. Ramo and J. R. Whinnery, *Fields and Waves in Modern Radio* (John Wiley and Sons, Inc., New York, 1944).

where a is the radius of the cavity, and $\rho=k_c r$. The last integral in (C-2) becomes

$$\int_0^{k_c a}\rho J_1'^2(\rho)d\rho=-\frac{J_1^2(k_c a)}{2}+k_c a J_0(k_c a)J_1(k_c a)+\frac{J_0^2(k_c a)}{2}-\frac{1}{2}-\frac{k_c^2 a^2}{2}[J_1^2(k_c a)-J_0(k_c a)J_2(k_c a)].\qquad\text{(C-3)}$$

The first integral in (C-2) cannot be integrated in a closed form, but the integrand can be expanded in a series and then integrated. For an upper limit of integration less than 2, only the first six terms are required for three-place accuracy. Thus

$$\int_0^{k_c a}\frac{1}{\rho}J_1^2(\rho)d\rho\doteq\left[\frac{\rho^2}{8}-\frac{\rho^4}{64}+\frac{5\rho^6}{4608}-\frac{7\rho^8}{147456}+\frac{\rho^{10}}{737{,}280}\right]_0^{k_c a}.\qquad\text{(C-4)}$$

Then the stored energy integral can be written

$$\frac{1}{2}\epsilon_0\int E^2 d\tau=\frac{1}{2}\epsilon_0 A_0^2\tau_0'=\frac{1}{2}\epsilon_0 E_0^2\tau_0$$

where τ_0' is found from the evaluation of (C-2) using expressions (C-3) and (C-4). Then $\tau_0=4\tau_0'$ because $A_0=2E_0$.

For an X band TE_{111} cavity, where $f_0=9000$ mc, $a=1.75\times10^{-2}$ meter, and $\lambda_g=4\times10^{-2}$ meter. Then $k_c a=1.85$, and the approximation condition is satisfied. For these numerical values, $\tau_0=2.28\times10^{-6}$ meter.[3]

Geometrical Factors Affecting the Contours of X-Ray Spectrometer Maxima. II. Factors Causing Broadening*

LEROY ALEXANDER
Department of Research in Chemical Physics, Mellon Institute, Pittsburgh, Pennsylvania
(Received July 25, 1949)

By applying Stokes' Fourier transform method for the analysis of diffraction maxima it is shown that the pure diffraction contour generated by a crystallite size distribution is apt to be approximated rather closely by the function $1/(1+k^2\phi^2)$. In the case of the x-ray spectrometer this pure diffraction contour is broadened significantly by the action of the following five geometrical factors: (I) the x-ray source width, (II) flat rather than curved sample surface, (III) vertical divergence of the x-ray beam, (IV) penetration of the sample by the beam, and (V) the receiving slit width.

The broadening of the pure diffraction contour due to the action of each of the five factors and the breadth of the final contour generated by the instrument can be deduced by employing the convolution approach suggested by Spencer. The effect of each instrumental factor is expressed by a convolution equation of the form

$$f_i(\phi)=\int_{-\infty}^{+\infty} W_i(\xi)f_{i-1}(\phi-\xi)d\xi,$$

in which ϕ is the angular displacement from twice the ideal Bragg angle, 2θ, f_{i-1} is the contour before the action of the ith geometrical factor, W_i is the form of the ith geometrical factor, and f_i is the contour after the action of W_i on f_{i-1}. Starting with a pure diffraction contour of the form $1/(1+k^2\phi^2)$, generalized broadening curves are derived for the effect of each of the five geometrical factors. Using these curves it is possible to predict the breadth of the final diffraction contour generated by the spectrometer from an initial contour of any breadth.

I. INTRODUCTION

IN addition to its theoretical interest this subject commands attention because of the very general employment of the diffraction contour width as a criterion of crystalline size. The problem is essentially that of deducing the nature of the pure diffraction contour free of instrumental contributions. Several investigators have suggested methods of attacking the problem,[1–6] the most directly applicable one being that of Stokes.[5]

In the present paper each of the geometrical factors causing significant broadening will be discussed separately, and it will be indicated how the net broadening can be estimated as a result of the combined action of the several factors. In the following paper (this issue) these findings are applied to the problem of more accurately determining the crystallite sizes of powders from contour broadening measurements.

When freed of instrumental effects, the breadth of a contour approaches zero for the case of crystallite dimensions greater than 1000A and for moderate Bragg angles (such as must be resorted to in conventional spectrometers). However, at Bragg angles close to 90° the pure contour breadth is measurable for crystallite dimensions as large as 2000A.

For a powder spectrometer in good adjustment the instrumental factors causing appreciable broadening are the following: (I) width of x-ray source (or "source" slit if two slits are used), (II) flat rather than curved sample surface, (III) vertical divergence, (IV) penetration of the sample by the beam, and (V) width of the receiving slit. In the older Norelco model in which Soller slits are not employed, factor (I) is usually dominant while (IV) and (V) may or may not be important. In the newer models vertical divergence is practically eliminated by the use of Soller slits, while at the same time very narrow sources are employed. For this case factor (I) becomes less important and (III) can be neglected.

II. THE MATHEMATICAL EXPRESSION OF INSTRUMENTAL EFFECTS

Spencer[2,3] has pointed out that the effect of an apparatus is to transform a function $f(x)$ into a function $F(x)$ which is directly observable. Assuming that the superposition theorem holds, $F(x)$ is the convolution (*Faltung*) of $f(x)$ and the weight function, $W(x)$, of the apparatus, *viz.*,

$$F(x)=\int_{-\infty}^{+\infty} W(u)f(x-u)du. \qquad (1)$$

This expression has also been derived by Jones[7] for the special case of Debye-Scherrer powder diffraction con-

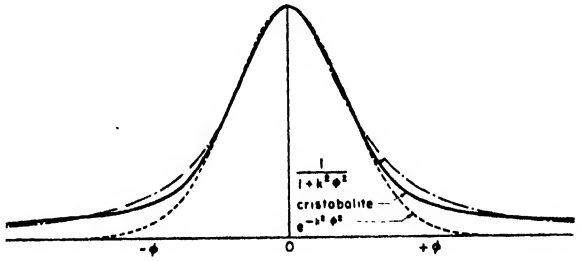


FIG. 1. Comparison of the pure diffraction contour of cristobalite with the functions $1/(1+k^2\phi^2)$ and $\exp(-k^2\phi^2)$.

* Presented in part before the American Society for X-Ray and Electron Diffraction at Cornell University, Ithaca, New York, on June 25, 1949. For Paper No. I, "Factors Causing Asymmetry," see J. App. Phys. **19**, 1068 (1948).

[1] Roy C. Spencer, Phys. Rev. **38**, 618 (1931).
[2] Roy C. Spencer, Phys. Rev. **55**, 239 (1939).
[3] Roy C. Spencer, J. App. Phys. **20**, 413 (1949).
[4] C. G. Shull, Phys. Rev. **70**, 679 (1946).
[5] A. R. Stokes, Proc. Phys. Soc. **61**, 382 (1948).
[6] A. Kochendorfer, Zeits. f. Krist. **105**, 393 (1944).
[7] F. W. Jones, Proc. Roy. Soc. **A166**, 16 (1938).

tours. The auxiliary variable u has the same dimensions as x and its introduction in Eq. (1) is a characteristic feature of the superposition theorem. The weight function, $W(x)$, is to be regarded as a function which expresses the sum total of the apparatus effects upon the pure function which is being measured. Alternately, the instrumental weight function may be split up into each of the several instrumental factors, which it may be more convenient to analyze separately. The ultimate response curve of the apparatus may then be obtained by the successive operation of each specific weight function upon the response curve generated by the action of the preceding specific function. Thus, for n instrumental factors and weight functions W_I, W_{II}, $W_{III}, \cdots W_n$, we would have the generation of the respective contours $f_1, f_2, f_3, \cdots, f_n = F$, according to the successive action of these weight functions:

$$\left.\begin{aligned} f_1(x) &= \int_{-\infty}^{+\infty} W_I(u) f(x-u)\,du, \\ f_2(x) &= \int_{-\infty}^{+\infty} W_{II}(u) f_1(x-u)\,du, \\ f_3(x) &= \int_{-\infty}^{+\infty} W_{III}(u) f_2(x-u)\,du, \\ f_n(x) = F(x) &= \int_{-\infty}^{+\infty} W_n(u) f_{n-1}(x-u)\,du. \end{aligned}\right\} \quad (2)$$

It should be borne in mind that in Eqs. (1) and (2), $f(x-u)$ is the pure function which we are seeking to measure and which was originally written simply as $f(x)$. It must be kept in mind that the symbol f denotes a specific function, regardless of the notation used for the independent variable (x or $x-u$).

The elucidation of contour broadening in an x-ray powder spectrometer, then, resolves itself into the twofold task of (a) deducing or assuming satisfactorily the nature of the pure diffraction contour, $f(x)$, and (b) determining the proper forms of the five instrumental weight functions W_I, W_{II}, W_{III}, W_{IV}, and W_V, corresponding to the five factors listed above which contribute to contour breadth. Once this has been done, Eqs. (2) can be employed to synthesize the final contour generated by the instrument starting from any original pure diffraction contour.

III. THE PURE DIFFRACTION CONTOUR

Case I. Crystallite Size > 1000A

For moderate Bragg angles the pure contour width is imperceptible and may be neglected. Hence, in Eq. (2) $f(x-u)$ is effectively a constant and $f_1(x) = W_I(x)$, so that for this case the first convolution equation drops out and we start with the action of W_{II}, the flat sample factor, on W_I, the x-ray source contour.

TABLE I. Pure diffraction contours of miscellaneous materials showing crystallite size broadening. The first five were analyzed by the Fourier method of Stokes. Degree of resemblance to $e^{-k^2\phi^2}$ or $1/(1+k^2\phi^2)$ is indicated as an approximate percentage.

No.	Material	Origin	Percentage resemblance to $e^{-k^2\phi^2}$	Percentage resemblance to $1/(1+k^2\phi^2)$
1	MgO	Decomposition of magnesium carbonate at 600°C.	20	80
2	CaF_2	Presumably precipitated from aqueous solution.	50	50
3	Cristobalite	Thermal treatment of diatomaceous earth (amorphous SiO_2).	20	80
4	Basic calcium phosphate	Precipitated from aqueous solution.	10	90
*5	Cold worked copper	Specific information not published.	20	80
**6	Nickel catalyst	Alkaline treatment of Raney alloy.	10	90

* This contour published by Stokes.
** This contour not analyzed by the Fourier method because its width was so great as to render the instrumental contributions negligible.

Case II. Crystallite Size < 1000A

The pure diffraction contour is a function of the mean crystallite shape[8] and of the size distribution. Obviously these factors will vary widely with the chemical constitution and origin of any particular powder sample. The investigations of Jones[7] tend to show that for a crystallite size distribution, the diffraction contour is given more closely by the function $1/(1+k^2x^2)$ than by a normal distribution curve ($e^{-k^2x^2}$). Stokes[5] has recently described a Fourier transform method for deducing the nature of the pure diffraction contour free of instrumental contributions. This method makes it possible for the first time to replace the somewhat speculative notions of the past with dependable information about the nature of the pure diffraction contour.

Using a Patterson-Tunell[9] set of strips and stencils for the rapid summation of Fourier series, the pure diffraction contours of a number of materials have been determined by the procedure of Stokes with the results summarized in Table I. In Fig. 1 the rather typical case of the cristobalite contour is portrayed. Its comparatively close coincidence with the function $1/(1+k^2\phi^2)$ is evident, and it will be noted from Table I that five of the six materials showed equally good or even better agreement with this function. Only in the case of CaF_2 was the agreement poorer, and even then the pure diffraction contour did not resemble $e^{-k^2\phi^2}$ more closely than $1/(1+k^2\phi^2)$. The most fundamental point of difference between these two mathematical functions is the much more gradual decrease in magnitude of the latter as ϕ becomes large. As suggested by Jones,[7] it is not surprising that this function should more accurately fit the case of the diffraction contour of a crystallite size distribution. There are theoretical reasons for $e^{-k^2\phi^2}$ being the better choice if all the

[8] A. R. Stokes and A. J. C. Wilson, Proc. Camb. Phil. Soc. 38, 313 (1942).
[9] A. L. Patterson and G. Tunell, Am. Mineralogist 27, 655 (1942).

crystallites are of the same size, whereas for a distribution of sizes the large crystallites will tend to sharpen the peak while the small ones will broaden the base of the contour.

In view of the above considerations together with the experimental data of Table I, there appears to be good justification for choosing the simple mathematical function $1/(1+k^2\phi^2)$ to approximate the pure diffraction contour generated by a typical crystallite size distribution.

Having thus decided on a suitable form for the pure contour function, $f(x-u)$, of Eqs. (2), we now proceed to evaluate each of the five weight functions corresponding to the five geometrical factors which act to modify it. In order to retain maximum generality in the following analysis, we shall discuss Case II, in which the pure diffraction contour possesses appreciable breadth. In the ensuing treatment as well as in the preceding three paragraphs the generalized parameter x of Eqs. (1) and (2) is replaced by the appropriate quantity ϕ, which is the angular displacement from the theoretical angle, 2θ, of a given reflection. Unless otherwise indicated the term *width*, or *breadth*, of a contour will refer to the width at half maximum intensity.

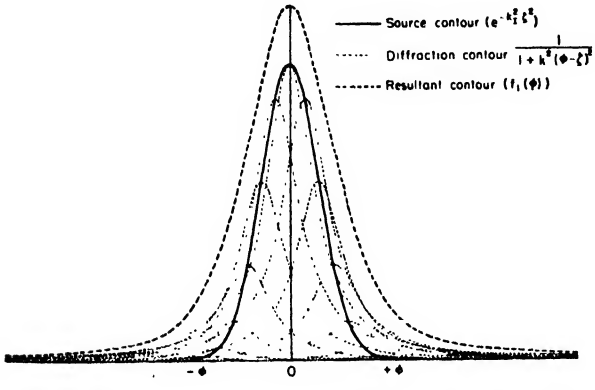


FIG. 2. Schematic representation of the convolution of the source contour and a pure diffraction contour:

$$f_1(\phi)=\int_{-\infty}^{+\infty}\frac{\exp(-k_I^2\zeta^2)d\zeta}{1+k^2(\phi-\zeta)^2}.$$

The case pictured is that of $w/w_I=0.85$ and $w_1/w=1.80$.

IV. EFFECT OF THE X-RAY SOURCE WIDTH (W_I)

In the past it has been customary to assume that the distribution of energy across the focus of the x-ray tube is either uniform or of the form $e^{-k^2\phi^2}$.[10] Pinhole photographs of tube foci usually reveal a more or less irregular intensity distribution which does not conform to either of these assumptions, although it commonly approaches the exponential state more or less closely. With a focusing powder spectrometer possessing a relatively wide source (of the order of $0.20°2\theta$), it is possible to observe the effective source contour with little distortion by scanning a diffraction contour produced by some material of large crystallite size in order to eliminate the pure contour breadth. At the same time it is necessary to keep the x-ray beam divergence small and to employ a narrow receiving slit. Under these conditions the factors other than source width become of minor importance in influencing the observed contour. Using a Norelco Geiger-Counter Spectrometer Type No. 12021, the copper and iron anode tubes used in the present investigation were thus shown to possess source configurations approaching the normal distribution curve, $e^{-k^2\phi^2}$, very closely. Such an intensity distribution will be assumed in the present paper, although it must be remembered that a radical departure from this type of configuration, which might well be found for some particular x-ray tubes, would lead to a corresponding modification of the analysis to follow.

We may write for the weight function of the source contour

$$W_I(\phi)=\exp(-k_I^2\phi^2), \tag{3}$$

in which the value of k_I determines the width of the contour and from which we have eliminated any amplitude coefficient since it has no bearing on the contour shape. The pure contour is assumed to be of the form

$$f(\phi)=1/(1+k^2\phi^2), \tag{4}$$

so that we obtain for the convolution of W_I and f

$$_1(\phi)=\int_{\infty}^{+\infty}W_I(\zeta)f(\phi-\zeta)d\zeta$$

$$=\int_{-\infty}^{+\infty}\frac{\exp(-k_I^2\zeta^2)d\zeta}{1+k^2(\phi-\zeta)^2}. \tag{5}$$

In terms of the widths of the W_I and f contours at half maximum intensity, w_I and w, we find for k_I and k respectively:

$$(W_I)_{\frac{1}{2}}=(W_I)_0/2=\tfrac{1}{2}=\exp(-k_I^2\phi_{\frac{1}{2}}^2)$$

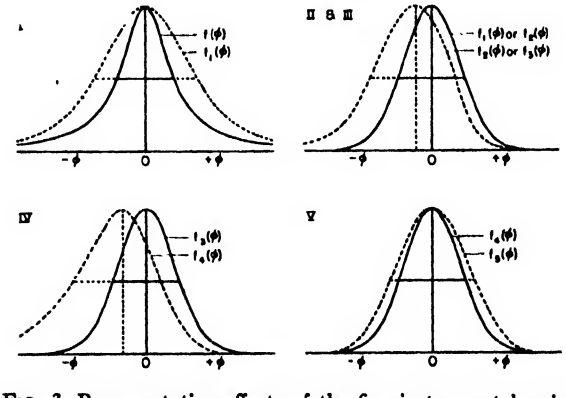


FIG. 3. Representative effects of the five instrumental weight functions upon a diffraction contour. The solid and dashed lines respectively depict the diffraction contour before and after the action of each weight function. The effects of the following functions are portrayed: (I) the x-ray source, (II) flat rather than curved sample surface, (III) vertical divergence, (IV) penetration of the sample by the beam, and (V) the receiving slit.

[10] A. Taylor and H. Sinclair, Proc. Phys. Soc. 57, 108 (1945).

or

$$k_I=(\log_e 2)^{\frac{1}{2}}/\phi_{\frac{1}{2}}=2(\log_e 2)^{\frac{1}{2}}/w_I=1.67/w_I, \quad (6)$$

and

$$f_{\frac{1}{2}}=f_0/2=\tfrac{1}{2}=1/(1+k^2\phi_{\frac{1}{2}}^2)$$

or

$$k=1/\phi_{\frac{1}{2}}=2/w. \quad (7)$$

Using the procedure described above for determining the approximate form of the source intensity distribution, it was found that w_I for the copper anode tube was about 0.20° (in $\Delta 2\theta$, or ϕ, units). This agreed substantially with a geometrical deduction of the angular source width from the manufacturer's specifications for the size of the focus. In the improved spectrometers with Soller slits and vertical focal spots, W_I is ordinarily very much smaller.

An attempt has been made in Fig. 2 to picture the meaning of the convolution equation, (5). The source contour acts as the envelope for an infinite number of diffraction contours (represented by a finite number in Fig. 2) all of the same shape but differing in amplitude. The resultant intensity at any point ϕ is the sum of the intensities of the various diffraction contours at this point. In Fig. 2 the amplitude of the resultant contour has been greatly scaled down to permit plotting it in the same diagram with the source and diffraction contours. It will be observed that the resultant contour is much broader than the original diffraction contour.

Equation (5) cannot be integrated readily, but sufficiently accurate solutions can be obtained by evaluating the equivalent summation at small intervals:

$$f_1(\phi)=\sum_{-\infty}^{+\infty}\frac{\exp(-k_I^2\zeta^2)}{1+k^2(\phi-\zeta)^2}\Delta\zeta. \quad (8)$$

The calculations are facilitated by graphing the denominator function for values of $(\phi-\zeta)$ and by utilizing standard tables of the normal distribution function, $(2\pi)^{-\frac{1}{2}}e^{-(\frac{1}{2})x^2}$, for obtaining the numerator. This is accomplished by converting the exponential function from $\exp(-k_I^2\zeta^2)$ to $\exp[-\frac{1}{2}(2^{\frac{1}{2}}k_I\zeta)^2]$, permitting the value of the numerator to be obtained as a function of the argument $2^{\frac{1}{2}}k_I\zeta$. When $w_I=0.20°$ the interval $\Delta\zeta$ should be about 0.01°. Figure 3(I) depicts the action of a source of breadth $w_I=0.20°$ in broadening a pure diffraction contour of breadth $w=0.20°$ to a resultant contour for which $w_1=0.327°$.

The relative broadening of the pure contour is a function only of the ratio of the breadth of the pure contour to the breadth of the x-ray source contour, w/w_I. The relative broadening is the ratio of the breadth of the contour after the action of the source factor, which we may denote by w_1, to its original breadth, w. Figure 4(I) shows this broadening ratio, w_1/w, as a function of w/w_I, the location of the curve having been precisely determined by the application of Eq. (8) to a number of pairs of values of w and w_I [and, therefore, also of k and k_I by Eqs. (6) and (7)].

V. EFFECT OF FLAT RATHER THAN CURVED SAMPLE SURFACE (W_{II})

From Fig. 5 it is seen that an x-ray beam of horizontal divergence angle α and impinging upon a flat sample surface at the angle θ (as in conventional spectrometers) will irradiate a length of sample which is approximately given by

$$2s=\alpha d/\sin\theta, \quad (9)$$

where d is the distance from source to sample. This formula is very accurate for the usual small values of α of not more than 2° or 3°. In actual practice the effective beam divergence at low Bragg angles is limited by the length of the sample, l, so that for $\sin\theta<\alpha d/l$ the effective divergence is $\alpha=(l/d)\sin\theta$. To simplify the present discussion, however, the sample surface will be assumed to be very large so that (9) holds at all values of θ. We shall first investigate the effect of sample flatness in broadening the focus of the diffracted rays at the receiving slit.

Figure 6 shows how this happens. The flat sample, S_1-S_2, is tangent to the focusing circle at its midpoint, S_0. A central ray FS_0, which impinges at the Bragg angle θ upon the surface, is diffracted (also at the Bragg angle θ) to the theoretical focus at G. But a non-central ray, FS', impinging at an angle $\theta+\Delta\theta$, is diffracted at

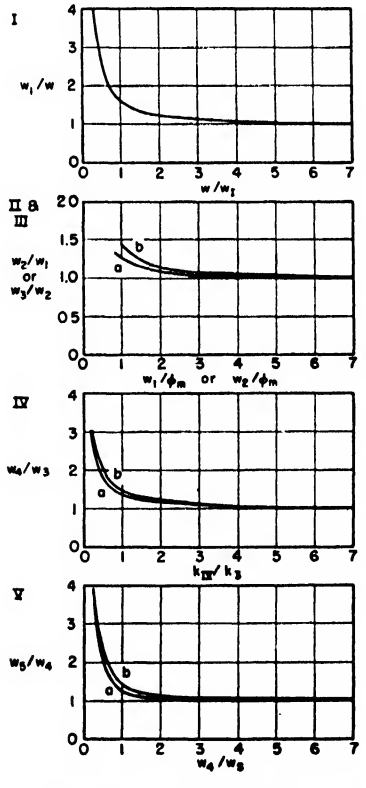


FIG. 4. Broadening ratios for the five instrumental weight functions, including (I) the x-ray source, (II) flat rather than curved sample surface, (III) vertical divergence, (IV) penetration of the sample by the beam, and (V) the receiving slit. Curves (a) and (b) apply respectively to $\exp(-k^2\phi^2)$ and $1/(1+k^2\phi^2)$ contours.

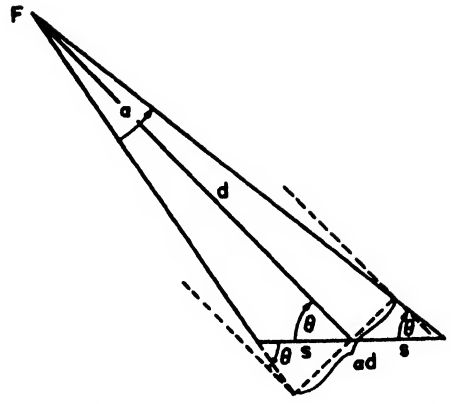


FIG. 5. Irradiation of a flat sample surface by a beam of divergence angle α.

the angle $\theta-\Delta\theta$, and arrives at some point G_1 which is displaced from the ideal focus by the linear interval GG_1. This displacement is proportional to z, the displacement of the point S' on the flat sample surface from the focusing circle. We now employ the approximate formula which expresses the length of the sagitta in terms of the circle radius r and the ordinate h (see Fig. 6, insert)

$$x=h^2/2r, \tag{10}$$

which is rather accurate for $h\ll r$. Applying this relationship to the situation of Fig. 6, we find that

$$z=h^2/2R, \tag{11}$$

where h is the distance S_0S' and R is the radius of the focusing circle. Since $R=d/2\sin\theta$,[11] this gives for z

$$z=h^2\sin\theta/d. \tag{12}$$

The displacement GG_1 from the ideal focus is then

$$GG_1=2z\cos(\theta+\Delta\theta)=(h^2/d)\sin2\theta, \tag{13}$$

or in terms of the angular deviation from the theoretical diffraction angle 2θ

$$\phi_{GG_1}=-(h^2/d^2)\sin2\theta \text{ radians}$$
$$=-57.3(h^2/d^2)\sin2\theta \text{ degrees}. \tag{14}$$

The minus sign indicates that the displacement ϕ is in the -2θ direction.

The limiting displacement of the diffracted rays from the ideal focus G corresponds to rays diffracted from the points S_1 and S_2, for which $h=s$. Then

$$\phi_m=-(s^2/d^2)\sin2\theta \text{ radians},$$

or, since $s\cong\alpha d/2\sin\theta$,

$$\phi_m=-\frac{\alpha^2}{4\sin^2\theta}\cdot 2\sin\theta\cos\theta$$
$$=-\tfrac{1}{2}\alpha^2\cot\theta \text{ radians}. \tag{15}$$

[11] L. Alexander, J. App. Phys. 19, 1068 (1948).

For ϕ and α in degrees this becomes

$$\phi_m=-\frac{\alpha^2\cot\theta}{114.6} \text{ degrees}. \tag{16}$$

This equation shows that the sharpness of focus (as affected by the flat sample factor) improves rapidly with decreasing beam divergence and increasing diffraction angle. If the finite length, l, of the sample surface is also taken into account, at angles for which $\sin\theta<\alpha d/l$ the limiting displacement of the diffracted rays from the ideal focus G is

$$\phi_m=-\tfrac{1}{2}(l^2/114.6d^2)\sin2\theta \text{ degrees}. \tag{17}$$

In practice, then, the sharpness of focus as a result of the flat sample effect is poorest (ϕ_m is a maximum) when $\sin\theta=\alpha d/l$ and it improves at both smaller and larger angles. We now need to determine the flat sample weight function, $W_{II}(\phi)$. From Eq. (13) or from an inspection of Fig. 6 it is seen that a displacement GG_1 (or ϕ_{GG_1}) is proportional to the sagitta variable z. But the intensity at a displacement ϕ from the theoretical focus is proportional to the energy diffracted in the interval Δz_ϕ (or $\Delta\phi_\phi$). This energy is proportional to the amount of sample surface diffracting into the increment Δz (or $\Delta\phi$), which is Δh_z (or Δh_ϕ). It follows, therefore, that the $W_{II}(\phi)$ function is proportional to $(\Delta h/\Delta z)$. But it has already been shown that we make negligible error in using the sagitta relationship for z [Eq. (11)], so that in the limit

$$W_{II}(\phi)\propto dh/dz=R/h=(R/2z)^{\frac{1}{2}}.$$

Since we are not interested in the absolute magnitude of the amplitude but only in the nature of the contour, and remembering that ϕ is proportional to z, we may dispense with a proportionality constant and write

$$W_{II}(\phi)=\phi^{-\frac{1}{2}}. \tag{18}$$

This function has the objection of an infinite ordinate at $\phi=0$, but the difficulty can be circumvented by evaluating the resulting convolution integral as a summation and omitting the zero term. For the convolution of $f_1(\phi)$ and $W_{II}(\phi)$ we then have

$$f_2(\phi)=\int_{-\infty}^{+\infty}W_{II}(\zeta)f_1(\phi-\zeta)d\zeta$$
$$=\int_{\phi_m}^{0}|\zeta|^{-\frac{1}{2}}\exp[-k_1^2(\phi-\zeta)^2]d\zeta, \tag{19}$$

assuming the f_1 contour to approximate an error curve. To evaluate $f_2(\phi)$ we go to the equivalent summation

$$f_2(\phi)=\sum_{\zeta=\phi_m}^{\zeta=\phi_1}|\zeta|^{-\frac{1}{2}}\exp[-k_1^2(\phi-\zeta)^2]\Delta\zeta \tag{20}$$

and sum at a relatively large number of ζ values (small

$\Delta\zeta$), omitting the $\zeta=0$ term. This omission has the effect of shifting the center of gravity of the contour slightly, but it will furnish a rather accurate contour shape provided $\Delta\zeta$ is kept small, e.g., of the order of $\phi_m/40$ or less. It will be noticed that ζ has been assigned to a range of negative rather than positive values. This is consistent with the sign convention of Eqs. (15) and (16) and is correct because the function W_{II} distorts a diffraction contour in the -2θ direction.

As in Part IV, the computations for actual experimental conditions are facilitated by graphing one function, this time $W_{II}(\zeta)=|\zeta|^{-\frac{1}{2}}$, and utilizing standard tables of the normal distribution function to evaluate $\exp[-k_1^2(\phi-\zeta)^2]$, after first writing the exponent as $-\frac{1}{2}[2^{\frac{1}{2}}k_1(\phi-\zeta)]^2$. For illustration we may outline briefly the computations for a rather extreme case of broadening, which portrays the nature of the results to be expected in a striking manner. The case is one of large beam divergence ($\alpha=1.41°$) and low Bragg angle ($\theta=5°$). From Eq. (16) we find for the greatest displacement of the diffracted rays from the theoretical focus:

$$\phi_m=-(\alpha^2\cot\theta)/114.6=-2\times11.43/114.6=-0.200°.$$

Assuming that the contour of the x-ray source is an error curve in form and that the width at half maximum intensity is 0.200° (there being no pure diffraction broadening assumed in this case), we find for k_1 by applying Eq. (6):

$$k_1=1.67/w_1=1.67/w_I=1.67/0.200=8.35 \text{ degrees}^{-1}.$$

When summed at intervals of $\Delta\zeta=0.005°$, Eq. (20) becomes

$$f_2(\phi)=\sum_{\zeta=-0.200°}^{\zeta=-0.005°}|\zeta|^{-\frac{1}{2}}\exp[-8.35^2(\phi-\zeta)^2]\times0.005°.$$

Dropping the constant factor 0.005° since it does not affect the nature of the contour, and converting the exponent into the standard error curve form, we have

$$f_2(\phi)=\sum_{\zeta=-0.200°}^{\zeta=-0.005°}|\zeta|^{-\frac{1}{2}}\exp[-\tfrac{1}{2}\{11.8(\phi-\zeta)\}^2]. \quad (21)$$

Figure 3(II) shows the $f_2(\phi)$ contour as computed with Eq. (21) and compares it with the original contour $f_1(\phi)=\exp(-8.35^2\phi^2)$. The peak has been broadened from $w_1=0.200°$ to $w_2=0.249°$, which means a broadening ratio of $w_2/w_1=1.245$. In addition the resulting peak is seen to be decidedly asymmetric and the center of gravity and maximum point have been shifted in the -2θ direction by appreciable amounts.

A consideration of Eq. (21) shows that the broadening ratio w_2/w_1 is a function only of the relative sizes of w_1 and ϕ_m. Accordingly if the $f_2(\phi)$ function is computed for a number of values of the ratio w_1/ϕ_m which cover all anticipated experimental situations, another curve of perfectly general applicability is obtained which is analogous to the curve of Fig. 4(I) developed in Part IV. This curve of w_2/w_1 as a function of w_1/ϕ_m is shown in Fig. 4(II,a). When the function $f_1(\phi-\zeta)$ of Eq. (19) has the form $1/[1+k_1^2(\phi-\zeta)^2]$, curve (b) is obtained. Curve (b) is more appropriate than (a) for cases in which the pure diffraction contour is much wider than the x-ray source contour. When w and w_I do not differ very greatly the broadening will be intermediate between the values predicted by curves (a) and (b).

For moderate values of the beam divergence, and Bragg angles of 10° or more, it is found that the broadening ratio seldom exceeds 1.02. This permits us to conclude that for ordinary experimental conditions the effect of flat sample surface is very minor.

VI. EFFECT OF VERTICAL DIVERGENCE OF THE X-RAY BEAM (W_{III})

In a preceding paper[11] the effect of vertical divergence and other third-dimensional properties of the spectrometer in producing asymmetry was discussed in detail. Referring to Fig. 2 of this reference, in the older spectrometers without Soller slit collimators the usual working value of the x-ray slit height, h_x, is considerably larger than the source height, h_f. Under these conditions we make but little error in assuming that the x-ray source has negligible height and that the centers of the diffraction halos in the plane $G'G$ (refer to Figs. 1 and 3 of the above reference) are uniformly distributed over a vertical interval

$$L=(h_x d/b)(1+\cos2\theta),$$

b and d being respectively the distances from the x-ray source to the x-ray slit and sample. Defining the vertical divergence angle β by h_x/b, this becomes

$$L=\beta d(1+\cos2\theta), \quad (22)$$

β being expressed in radians.

As a first approximation we may assume negligible receiving slit height. In that case, referring to Fig. 3 of the above reference, the maximum displacement of a diffracted ray from the ideal receiver focus is given by the sagitta of the arc of height $\frac{1}{2}L$ and radius r. Em-

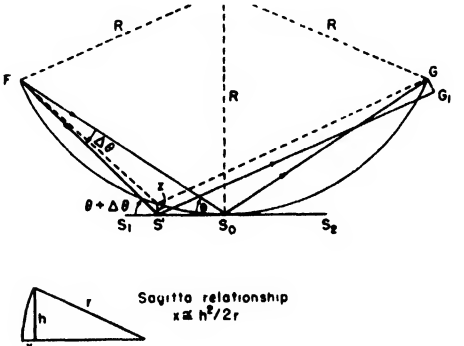


FIG. 6. Displacement of a diffracted ray from the theoretical focus G because of the use of a flat rather than curved sample surface.

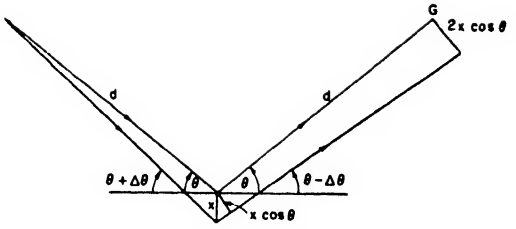


FIG. 7. Displacement of a diffracted ray from the theoretical focus G due to penetration of the sample to a depth x.

ploying the approximate sagitta formula as was done in Part V, we find for this displacement

$$x\ (\text{cm}) = (L/2)^2/2r = L^2/8r; \qquad (23)$$

but from Fig. 1 of reference 11 it can be seen that a displacement x in the direction GG' equals an angular displacement ϕ tangent to the spectrometer arc of

$$\phi = x\ (\text{cm}) \times (\cos 2\theta)/d \text{ radians}.$$

Combining this with Eqs. (22) and (23) and noting that $r = d \sin 2\theta$, we obtain

$$\phi_m = -(\beta^2/8)(1+\cos 2\theta)^2 \cot 2\theta \text{ radians}, \qquad (24)$$

the negative sign indicating that the displacement from the theoretical angle 2θ is in the negative direction. For ϕ_m and β in degrees Eq. (24) becomes

$$\phi_m = -(\beta^2/8 \times 57.3)(1+\cos 2\theta)^2 \cot 2\theta \text{ degrees}. \qquad (25)$$

The preceding analysis demonstrates that, as in Part V, we are again dealing with an instrumental weight function, this time $W_{III}(\phi)$, which varies as the inverse square root of the displacement, ϕ, from the ideal angle of deviation, 2θ. This follows in a manner parallel to that involved in the development of Eq. (18) since the intensity corresponding to a displacement ϕ from the theoretical focus is proportional to the energy diffracted in the interval Δx_ϕ [or $(\Delta\phi)_\phi$], and this energy is proportional to $(\Delta L)_\phi$. Thereupon in the limit

$$W_{III}(\phi) \propto \frac{dL}{dx} = \left(\frac{2r}{x}\right)^{\frac{1}{2}},$$

and, neglecting the constant factor as in Part V, and remembering that x is proportional to ϕ, we again arrive at the weight function,

$$W_{III}(\phi) = \phi^{-\frac{1}{2}}. \qquad (26)$$

Assuming $f_2(\phi)$ to approximate an error curve, we then find for the convolution, $f_3(\phi)$, of $W_{III}(\phi)$ and $f_2(\phi)$

$$f_3(\phi) = \int_{-\infty}^{+\infty} W_{III}(\zeta) f_2(\phi-\zeta)\, d\zeta,$$

or

$$f_3(\phi) = \sum_{\zeta=\phi_m}^{\zeta=\phi_1} |\zeta|^{-\frac{1}{2}} \exp[-k_2^2(\phi-\zeta)^2]\Delta\zeta, \qquad (27)$$

which is identical in form to Eq. (20).

It is evident that the remarks made in connection with Eq. (20) apply here, and also that curves (a) and (b) of Fig. 4 (II and III) apply directly in the present case for w_3/w_2 as a function of w_2/ϕ_m. The only difference is that Eq. (25) instead of Eq. (16) is to be used to determine ϕ_m for the particular experimental conditions under consideration.

For Bragg angles of 10° or more and for vertical divergence angles less than 2.4° it is found that the broadening ratio is seldom greater than 1.10 and more commonly it is of the order of 1.05. Evidently vertical divergence is much more important than flat sample surface in producing contour broadening.

VII. EFFECT OF PENETRATION OF THE SAMPLE BY THE X-RAY BEAM (W_{IV})

Consider a ray incident upon the sample at the angle $\theta+\Delta\theta$ and diffracted from a crystallite at a depth x (Fig. 7). If $\Delta\theta$ is neglected because it is very small, the intensity of the emergent diffracted ray, I, relative to that of a ray diffracted under conditions of non-absorption, I_0, is

$$I/I_0 = \exp(-2\mu x/\sin\theta), \qquad (28)$$

μ being the linear absorption coefficient of the powder. The linear displacement of this diffracted ray from the theoretical focus at the receiver is evidently $2x\cos\theta$, and the corresponding angular displacement, ϕ_x, is

$$\phi_x = -(114.6x\ \cos\theta)/d \text{ degrees}, \qquad (29)$$

the negative sign denoting a displacement in the -2θ direction. Equation (28), then, is an expression of the instrumental weight function, W_{IV}, due to penetration of a thick powder sample by the x-ray beam, and when this is combined with Eq. (29) we obtain W_{IV} expressed as a function of ϕ:

$$W_{IV}(\phi) = e^{k_{IV}\phi}, \qquad (30)$$

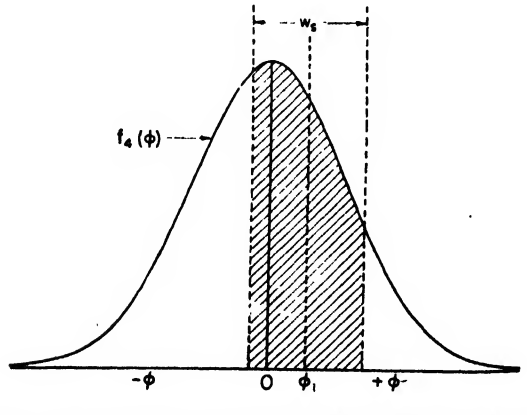


FIG. 8. Representation of a receiving slit of width w_s scanning a diffraction contour $f_4(\phi)$. The resulting intensity is proportional to the area of the contour intercepted (shaded zone).

in which

$$k_{IV}=4\mu d/114.6\sin 2\theta, \tag{31}$$

and ϕ varies between the limits 0 and $-\infty$.

Assuming first that $f_3(\phi)$ approximates an error curve (i.e., that the pure diffraction broadening is slight and that the preceding convolution sequence has not greatly modified the original error curve contour of the x-ray source), we obtain for the convolution, $f_4(\phi)$, of $W_{IV}(\phi)$ and $f_3(\phi)$

$$f_4(\phi)=\int_{-\infty}^{+\infty} W_{IV}(\zeta)f_3(\phi-\zeta)d\zeta,$$

or

$$f_4(\phi)=\int_0^{-\infty} e^{k_{IV}\zeta}\exp[-k_3^2(\phi-\zeta)^2]d\zeta. \tag{32}$$

By combining the two exponential factors and removing a constant factor from the integral, Eq. (32) can be put in the form

$$f_4(\phi)=\exp\left[k_{IV}\left(\phi+\frac{k_{IV}}{4k_3^2}\right)\right]\times\int_0^{-\infty}\exp\left[-\left(k_3\zeta-\frac{k_{IV}+2k_3^2\phi}{2k_3}\right)^2\right]d\zeta.$$

A change of variable from ζ to

$$u=2^{\frac{1}{2}}\left(k_3\zeta-\frac{k_{IV}+2k_3^2\phi}{2k_3}\right)$$

then leads to the result

$$f_4(\phi)=\exp\left[\left(k_{IV}\phi+\frac{k_{IV}^2}{4k_3^2}\right)\Big/2^{\frac{1}{2}}k_3\right]\times\int_l^{-\infty}\exp[-(u^2/2)]du, \tag{33}$$

in which the lower limit of the integral, l, is given by

$$l=(k_{IV}/2^{\frac{1}{2}}k_3)+2^{\frac{1}{2}}k_3\phi. \tag{34}$$

The expression for the convolution $f_4(\phi)$ has thus been reduced to the state of an integral (Eq. (33)) which can be evaluated directly by reference to tables of the normal curve of error. For any particular sample material of linear absorption coefficient μ and for a given spectrometer setting 2θ the value of k_{IV} can be calculated with Eq. (31), while k_3 depends upon the width at half maximum intensity of the error curve contour $f_3(\phi)$ according to a relationship identical with Eq. (6), or

$$k_3=1.67/w_3. \tag{35}$$

Figure 3(IV) depicts the effect of sample penetration upon a contour of the form $f_3(\phi)=\exp[-(k_3^2\phi^2)]$ for the particular case of $k_{IV}/k_3=1$. This case was chosen because it is typical of experimental conditions under which the effects of sample penetration are conspicuous. The given values of k are satisfied by either of the following sets of experimental conditions, the first corresponding to the old Norelco spectrometer with wide source ($w_I=0.20°$) and the second to a spectrometer with a relatively narrow source ($w_I=0.07°$):

I. Width of source contour: $w_I=0.20°$
Bragg angle : $\theta=15°$
Linear absorption coefficient of powder : $\mu=9$ or $\begin{cases}\theta\to 45°\\ \mu=18\end{cases}$

(CuKα radiation and powders consisting of metal-organic compounds or mixtures of organic compounds and common siliceous minerals)

II. Width of source contour: $w_I=0.07°$
Bragg angle : $\theta=15°$
Linear absorption coefficient of powder : $\mu=27$ or $\begin{cases}\theta\to 45°\\ \mu=54\end{cases}$

(CuKα radiation and powders prepared from common siliceous minerals)

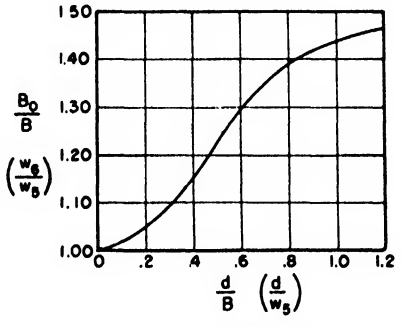


FIG. 9. Inverse of Jones' curve for correcting an observed contour breadth for Kα doublet broadening. B_0=observed breadth, B=breadth corrected for Kα doublet, d=angular separation of α_1 and α_2.

The broadening ratio for this case, w_4/w_3, is a function only of the ratio of the contour constants, k_{IV}/k_3. Curve (a) of Fig. 4(IV) depicts this function for the case just discussed wherein the $f_3(\phi)$ contour is given by $\exp[-(k_3^2\phi^2)]$. When $f_3(\phi)$ is given by $1/(1+k_3^2\phi^2)$, a condition approached by materials giving considerable pure diffraction broadening, the convolution integral cannot be evaluated directly as in the case of Eq. (32) but must be treated as the equivalent summation in the manner described in earlier sections. In this case curve (b) of Fig. 4(IV) is obtained.

If the absorption coefficient of the sample is low and if θ is not very small, sample penetration can evidently be an important source of contour broadening. Thus, in the example cited above for a powder of large crystallite size the broadening ratio was found to be 1.40. In the newer spectrometers with narrow sources this instrumental factor is relatively much more important than in the old Norelco instrument. For example, if $k_{IV}/k_3=2.86$ for $w_I=0.20°$, it will be decreased to 1.00 for $w_I=0.07°$, and reference to curve (a) of Fig. 4(IV) shows that this increases the broadening ratio from 1.14

TABLE II. Summarizing table of the effects of the instrumental weight functions upon the diffraction contour.*

Instrumental weight function	Contour function Initial	Contour function Final	Contour width Initial	Contour width Final	Form of convolution expression General	Form of convolution expression Spectrometer case	Constants Instrumental	Constants Diffraction contour	Integral limits
W_I Source	$f(\phi)$	$f_1(\phi)$	w	w_1	$f_1(\phi)=\int_{-\infty}^{+\infty} W_I(\zeta)f(\phi-\zeta)d\zeta$	$\int_{-\infty}^{+\infty}\frac{\exp(-k_I^2\zeta^2)d\zeta}{1+k^2(\phi-\zeta)^2}$	$k_I=\frac{1.67}{w_I}$	$k=\frac{2}{w}$	
W_{II} Flat sample	$f_1(\phi)$	$f_2(\phi)$	w_1	w_2	$f_2(\phi)=\int_{-\infty}^{+\infty} W_{II}(\zeta)f_1(\phi-\zeta)d\zeta$	(A) $\int_{\phi_m}^{0}\lvert\zeta\rvert^{-\frac{1}{2}}\exp[-k_1^2(\phi-\zeta)^2]d\zeta$ (B) $\int_{\phi_m}^{0}\frac{\lvert\zeta\rvert^{-\frac{1}{2}}d\zeta}{1+k_1^2(\phi-\zeta)^2}$		$k_1=\frac{1.67}{w_1}$ $k_1=\frac{2}{w_1}$	$\phi_m=-\frac{\alpha^2\cot\theta}{114.6}$
W_{III} Vertical divergence	$f_2(\phi)$	$f_3(\phi)$	w_2	w_3	$f_3(\phi)=\int_{-\infty}^{+\infty} W_{III}(\zeta)f_2(\phi-\zeta)d\zeta$	(A) $\int_{\phi_m}^{0}\lvert\zeta\rvert^{-\frac{1}{2}}\exp[-k_2^2(\phi-\zeta)^2]d\zeta$ (B) $\int_{\phi_m}^{0}\frac{\lvert\zeta\rvert^{-\frac{1}{2}}d\zeta}{1+k_2^2(\phi-\zeta)^2}$		$k_2=\frac{1.67}{w_2}$ $k_2=\frac{2}{w_2}$	$\phi_m=-\frac{\beta^2}{8\times57.3}(1+\cos2\theta)^2\cdot\cot2\theta$
W_{IV} Sample penetration	$f_3(\phi)$	$f_4(\phi)$	w_3	w_4	$f_4(\phi)=\int_{-\infty}^{+\infty} W_{IV}(\zeta)f_3(\phi-\zeta)d\zeta$	(A) $\int_{-\infty}^{0}e^{k_{IV}\zeta}\exp[-k_3^2(\phi-\zeta)^2]d\zeta$ (B) $\int_{-\infty}^{0}\frac{e^{k_{IV}\zeta}d\zeta}{1+k_3^2(\phi-\zeta)^2}$	$k_{IV}=\frac{4\mu d}{114.6\sin2\theta}$	$k_3=\frac{1.67}{w_3}$ $k_3=\frac{2}{w_3}$	
W_V Receiving slit	$f_4(\phi)$	$f_5(\phi)$	w_4	w_5	$f_5(\phi)=\int_{-\infty}^{+\infty} W_V(\zeta)f_4(\phi-\zeta)d\zeta$	(A) $\int_{\phi-w_s/2}^{\phi+w_s/2}\exp[-k_4^2\zeta^2]d\zeta$ (B) $\int_{\phi-w_s/2}^{\phi+w_s/2}\frac{d\zeta}{1+k_4^2\zeta^2}$		$k_4=\frac{1.67}{w_4}$ $k_4=\frac{2}{w_4}$	
W_{VI} Kα doublet	$f_5(\phi)$	$f_6(\phi)$	w_5	w_6					

* ϕ = angular displacement from the theoretical diffraction angle, 2θ.
$f(\phi)$ = pure diffraction contour.
w = width at half maximum intensity of the pure diffraction contour.
(A) Form of the convolution expression for a $\exp[-(k^2\phi^2)]$ contour.
(B) Form of the convolution expression for a $1/(1+k^2\phi^2)$ contour.

to 1.40. When powders of moderate or large absorption coefficients are used, the broadening ratio for the old Norelco spectrometer using CuKα radiation is seldom in excess of 1.20 and it often is less than 1.10. Even so, this instrumental factor is more important, in general, than the flat sample or vertical divergence factor in causing contour broadening.

The present analysis of sample penetration has presupposed the use of an infinitely thick sample. This condition is effectively realized under spectrometer conditions if the sample thickness, t, is great enough to satisfy the condition[12]

$$t \geq \frac{3.2}{\mu} \cdot \frac{\rho}{\rho'} \sin\theta, \tag{36}$$

in which μ and ρ are respectively the linear absorption coefficient and density of the solid material composing the powder, and ρ' is the density of the powder including interstices. The contour distortion due to sample penetration can be minimized if necessary in special cases, but at the expense of reduced intensity, by mounting a very thin layer of powder on a suitable flat substrate.

VIII. EFFECT OF RECEIVING SLIT WIDTH (W_V)

Figure 8 illustrates diagrammatically the effect of the receiving slit upon the contour being scanned ($f_4(\phi)$ in our present notation). It is assumed that the sensitivity of all points within the slit area is uniform. When the center of the slit is set at an angle of deviation ϕ_1 from the peak maximum (ideally 2θ), the slit intercepts a portion of the contour extending between the limits $\phi_1 - w_s/2$ and $\phi_1 + w_s/2$, w_s being the angular width of the slit. The recorded intensity, $f_5(\phi_1)$, corresponding to this position is evidently proportional to the area intercepted. Hence, we can write directly

$$f_5(\phi) = K \int_{\phi - w_s/2}^{\phi + w_s/2} \exp[-(k_4^2\zeta^2)]d\zeta, \tag{37}$$

assuming $f_4(\phi)$ to be an error curve in form. If we assume it to be more closely represented by the function $1/(1+k^2\phi^2)$,

$$f_5(\phi) = K \int_{\phi - w_s/2}^{\phi + w_s/2} \frac{d\zeta}{1+k_4^2\zeta^2}. \tag{38}$$

In Eqs. (37) and (38) the weight function $W_V(\phi)$ may be interpreted as a constant, K, between the limits $\zeta = \phi - w_s/2$ and $\phi + w_s/2$. Equation (37) can be readily evaluated by transforming the exponent to $\frac{1}{2}(2^{\frac{1}{2}}k_4\zeta)^2$ and referring to tabulated values of the area of the standard error curve between definite limits. The proper value of k_4 is deduced, as explained earlier, by Eq. (6) or (7) depending upon which form of contour, $f_4(\phi)$, is assumed. The angular equivalent of the linear slit width, w_s (cm), is given by

$$w_s \text{ (degrees)} = w_s \text{ (cm)} \times 57.3/d,$$

d being the sample-to-receiving slit distance. This angular value of w_s is used in determining the integral limits $\phi + w_s/2$ and $\phi - w_s/2$.

Figure 3(V) shows the effect of a receiving slit of width $w_s = 0.045$ cm $= 0.20°$ upon an error curve diffraction contour for which $w_4 = 0.20°$ (corresponding to $k_4 = 1.67/w_4 = 8.35$). The width of the resulting contour, $f_5(\phi)$, is $w_5 = 0.25°$, which corresponds to a broadening ratio of $w_5/w_4 = 0.25/0.20 = 1.25$. Figure 4(V) shows plots of the broadening ratio w_5/w_4 for the two cases of (a) $f_4(\phi) = \exp[-(k_4^2\phi^2)]$ and (b) $f_4(\phi) = 1/(1+k_4^2\phi^2)$.

TABLE III. Application of the method of convolution analysis to four representative cases:* Case (a) $D > 1000$ A, $w = 0°$, $w_I = 0.20°$. Case (b) $D > 1000$ A, $w = 0°$, $w_I = 0.07°$, Soller slits. Case (c) $D = 300$ A, $w = 0.305°$ by Scherrer formula, $w_I = 0.20°$. Case (d) $D = 300$ A, $w = 0.305°$ by Scherrer formula, $w_I = 0.07°$, Soller slits.

Instrumental weight function	Quantity tabulated	Values of the quantities listed in column 2 Case (a)	Case (b)	Case (c)	Case (d)
I. Source	w	0	0	0.305°	0.305°
	w/w_I	0	0	1.53	4.36
	w_1/w	—	-	1.34	1.05
	w_1	0.200°	0.070°	0.409°	0.320°
II. Flat Sample	w_1/ϕ_m	5.89	2.06	12.0	9.4
	w_2/w_1	1.01	1.06	1.00	1.00
($\phi_m = -.034°$)	w_2	0.202°	0.074°	0.409°	0.320°
III. Vertical Divergence	w_2/ϕ_m	2.77	-	5.6	--
	w_3/w_2	1.033	—	1.016	—
($\phi_m = -.073°$)	w_3	0.209°	0.074°	0.416°	0.320°
IV. Sample Penetration	k_{IV}/k_3	4.55	1.61	8.2	5.81
	w_4/w_3	1.035	1.20	1.02	1.05
	w_4	0.216°	0.089°	0.424°	0.336°
V. Receiving Slit	w_4/w_s	2.46	1.01	4.82	3.82
	w_5/w_4	1.045	1.25	1.015	1.035
	w_5	0.226°	0.111°	0.430°	0.348°
VI. Kα Doublet	d/w_5	0.354	0.721	0.186	0.230
	w_6/w_5	1.135	1.36	1.04	1.06
	w_6	0.254	0.151°	0.447°	0.369°

* D refers to the crystallite size. Other conditions assumed are: $\mu = 40$, $2\theta = 30°$, $\lambda = 1.542$A (CuKα), $w_s = 0.088°$.

IX. EFFECT OF UNRESOLVED Kα DOUBLET (W_{VI})

For Bragg angles to small to resolve the Kα doublet, the observed contour breadth will also be increased as a result of the superposition of the α_1 and α_2 contours. The width of either contour will be w_5 as a result of the action of the five instrumental factors (see Part VIII). The broadening due to an unresolved doublet has been thoroughly analyzed by Jones,[7] who developed a curve for correcting for this effect. For our purpose Jones' plot of B/B_0 as a function of d/B_0 can be more readily used if transformed to a plot of B_0/B as a function of d/B (w_6/w_5 against d/w_5 in the present notation). This curve is shown in Fig. 9. Here w_5 and w_6 represent respectively the contour widths before and after

[12] L. Alexander and H. P. Klug, Anal. Chem. 20, 886 (1948).

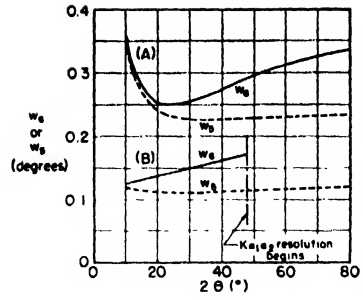


FIG. 10. Predicted dependence of contour breadth upon 2θ for spectrometers with x-ray source widths of (A) 0.20° and (B) 0.07° 2θ. w_b = width of a contour not resolving the Kα doublet; w_s = width of a resolved line of the doublet. Conditions assumed are: $\mu = 40$, $w_s = 0.088°$, crystallite size greater than 1000 A.

broadening due to the unresolved doublet, and d is the angular separation of the Kα doublet characteristic of the particular radiation and diffraction angle involved (see *Internationale Tabellen zur Bestimmung von Kristallstrukturen*, Vol. II, p. 588 ff.).

X. APPLICATION TO REPRESENTATIVE EXPERIMENTAL CASES

Table II summarizes the results of the analyses presented in the preceding sections. The five instrumental weight functions are presented in the order in which it is logical to apply them when making use of the broadening curves given in Fig. 4. Although not strictly an instrumental factor, the effect of Kα doublet broadening is included (W_{VI}) because of the necessity of taking it into account in arriving at the actually observed contour breadth at low angles. Table III illustrates with numerical data the application of the analysis to four representative experimental cases. In each case it is assumed that $\mu = 40$ (moderate absorbing power for x-rays), $2\theta = 30°$, $\lambda = 1.542$A (CuKα radiation), and the angular width of the receiving slit, w_s, is 0.088° (a relatively narrow slit). The four specific cases carried through in the four columns of the table are:

(a) Crystallite size > 1000A, $w \to 0°$, $w_I = 0.20°$ (older spectrometer).

(b) Crystallite size > 1000A, $w \to 0°$, $w_I = 0.07°$ (newer spectrometers with narrow source and Soller slits).

(c) Crystallite size 300A, $w = 0.305°$ by Scherrer formula, $w_I = 0.20°$.

(d) Crystallite size 300A, $w = 0.305°$ by Scherrer formula, $w_I = 0.07°$.

A few important observations can be made. With a relatively wide x-ray source contour (Cases (a) and (c)) and assuming the use of a narrow receiving slit, the dominant instrumental factor causing broadening is the source width. All the other instrumental factors are inappreciable or nearly so, and the Kα doublet becomes significant only for the larger crystallite sizes. In particular, the flat sample and vertical divergence effects are very slight. These become greater at small diffraction angles since the contour widening increases with the absolute magnitude of ϕ_m for both effects (see Eqs. (17) and (25)), although flat sample broadening diminishes again when $\sin\theta$ becomes less than $\alpha d/l$ (see Part V).

The same remarks apply to the case of a narrow source except that the source is relatively less significant and the other factors more significant as causes of contour broadening. Except at low Bragg angles the flat sample and vertical divergence effects remain very minor. Penetration of the sample and Kα doublet separation both become important at the angle considered ($2\theta = 30°$), particularly for contours displaying little crystallite size broadening.

In Fig. 10 the convolution analysis is applied to the problem of predicting the observed contour breadth as a function of 2θ (A) for the old Norelco spectrometer with a source width of 0.20° 2θ and (B) for newer spectrometers with an assumed source width of 0.07° 2θ and Soller slits to eliminate vertical divergence. The curves are plotted for the typical case of large crystallite size, moderately absorbing sample ($\mu = 40$), and a relatively narrow receiving slit width of 0.088° 2θ. It is also assumed that the length of sample is sufficient to receive the entire beam at the lowest angle indicated ($2\theta = 10°$). It will be noticed that resolution of the Kα doublet is obtained below 50° for the newer design whereas in the older design the doublet remains unresolved through the entire angular range. The failure of Curve B to make a sharp upturn below 20° such as occurs for Curve A is largely due to the absence of vertical divergence.

ACKNOWLEDGMENT

The author is appreciative of the interest and encouragement contributed by Dr. Harold P. Klug during the conduct of this investigation. He wishes to thank Mr. Douglas Pitman for valuable help in obtaining data and in performing calculations. He is also indebted to Dr. John R. Bowman for aid in evaluating the integral of Eq. (32).

Determination of Crystallite Size with the X-Ray Spectrometer*

LEROY ALEXANDER AND HAROLD P. KLUG
Department of Research in Chemical Physics, Mellon Institute, Pittsburgh, Pennsylvania
(Received July 25, 1949)

The accuracy of the Scherrer crystallite size equation is limited in part by the uncertainty in β, the experimentally deduced pure diffraction broadening. Currently used procedures for estimating β from the observed breadth of a Debye-Scherrer line are not, in general, applicable to the x-ray spectrometer.

By making use of a scheme of convolution analysis for analyzing the effect of geometrical factors in broadening the pure diffraction contour, a correction curve is developed for determining β from the experimentally measured line breadths b and B (Jones' notation). The degree of reliability of this correction procedure is ascertained by applying Stokes' direct Fourier transform procedure for determining the form of the pure diffraction contour free of instrumental effects.

Suggestive procedures are given for crystallite size determination with the x-ray spectrometer in different size ranges, and several examples are described.

I. INTRODUCTION

THE accuracy of the familiar Scherrer equation,

$$D=(K\lambda)/(\beta\cos\theta), \tag{1}$$

is limited by the uncertainties in K, the crystallite shape factor, and β, the pure diffraction broadening. Theoretical work by Stokes and Wilson[1] has elucidated the relationship between K and the crystal shape to a degree not previously attained. However, for the usual applications to crystallite size measurement a large degree of indeterminancy in K, as much as 20 percent or even more, remains, and it is customary to let $K=1$. To gain optimum accuracy it remains to reduce the uncertainty in β as much as possible, preferably to a value much less than 20 percent.

Considerable effort has been expended by a number of investigators in evolving expressions for readily deducing β from B, the observed breadth, for the case of Debye-Scherrer lines. The most widely applied are the procedures due to Warren[2] and Jones.[3] These published correction curves have also been applied to the analysis of spectrometer maxima, although no one has specifically demonstrated their applicability to the particular conditions governing the generation of spectrometer contours. The formula of Warren,

$$\beta^2=B^2-b^2, \tag{2}$$

where b is the observed breadth of a line produced by crystallites of sufficient size to give no diffraction broadening, applies only to the case in which the β and b contours are both true "error" curves of the form $e^{-k^2x^2}$. Similarly Jones' two correction curves (a) and (b) apply only to the cases where the β contour has the respective forms $e^{-k^2x^2}$ and $1/(1+k^2x^2)$. The form of the b contour assumed by Jones was experimentally obtained by microphotometering a Debye-Scherrer line at a Bragg angle large enough to clearly resolve the Kα doublet.

Recently Shull[4] and Stokes[5] have developed Fourier transform procedures for deducing the β contour from observed b and B contours. The method of Stokes is the more generally applicable but it requires a means for the rapid summation of Fourier series. More recently one of the present authors[6] has shown how the convolution, or *Faltung*, theorem can be used to synthesize the form of a b or B contour generated by an x-ray spectrometer starting with a pure diffraction contour of known dimensions. In the light of these developments it seemed that this was a propitious time to examine the problem of crystallite size determination with the x-ray spectrometer with a view to improving the accuracy with which the width of the pure diffraction contour can be deduced free of the various instrumental broadening effects.

Since most of the commercial spectrometers in use today are of the older Norelco design which is limited to the θ range from 0° to 45°, it is felt desirable to concentrate the present analysis on crystallite size determinations which can be carried out in this angular range. Hence, we shall consider size measurements mainly in the range 0 to 500A and particularly those between about 250 and 500A, where the errors due to improper correction of the instrumental contributions may be very large. With the older type of spectrometer, having a wide x-ray source contour and limited to angles below $\theta=45°$, it is futile to attempt size measurements for crystallites much larger than 500A because the mandatory precision of the experimental measurements becomes too great for practical work.

II. COMPUTATION OF A CORRECTION CURVE FOR CRYSTALLITE SIZE DETERMINATION WITH THE X-RAY SPECTROMETER

In reference 6 the broadening effects of the five significant instrumental factors, or weight functions,

* Presented in part before the American Society for X-Ray and Electron Diffraction at Cornell University, Ithaca, New York, on June 25, 1949.

[1] A. R. Stokes and A. J. C. Wilson, Proc. Camb. Phil. Soc. 38, 313 (1942); *ibid.* 40, 197 (1944).
[2] B. E. Warren, J. App. Phys. 12, 375 (1941).
[3] F. W. Jones, Proc. Roy. Soc. A166, 16 (1938).
[4] C. G. Shull, Phys. Rev. 70, 679 (1946).
[5] A. R. Stokes, Proc. Phys. Soc. 61, 382 (1948).
[6] L. Alexander, J. App. Phys. 21, 126 (1950).

TABLE I. Data utilized in the preparation of the spectrometer correction curve. $\theta=30°$, $w_l=0.20°$, $w_s=0.088°$. By convolution analysis $b=0.210°$.

D(A)	β(°) (from Scherrer formula)	B(°) (from convolution theory)	b/B	β/B
∞	0	0.210(b)	—	—
1000	0.102	0.272	0.773	0.375
700	0.1455	0.300	0.700	0.485
500	0.204	0.341	0.616	0.598
300	0.341	0.444	0.473	0.769
200	0.510	0.596	0.352	0.856
100	1.020	1.061	0.198	0.962
50	2.040	2.040	0.103	1.000

upon a pure diffraction contour were investigated. A scheme of convolution analysis was employed to develop generalized broadening curves for each of the factors, and it was shown that by successively employing each of these curves the width of the diffraction contour generated by the spectrometer could be predicted, starting with a pure diffraction contour of any assumed width.

In reference 6 it was pointed out that with a proper choice of the experimental conditions three of the five instrumental factors causing contour broadening become negligible and only the effects of (I) the x-ray source width and (V) the receiving slit width need be considered. The requisite conditions are that the linear absorption coefficient of the powder for the x-ray beam be not too small ($\mu \geqq 40$) and that the diffraction angle 2θ be 20° or greater. The first condition is met in the case of CuKα radiation for nearly all materials except organic and metal-organic compounds, and the second condition can nearly always be satisfied by the proper choice of a diffraction line. It will accordingly be assumed in the development to follow that these simplifying conditions obtain.

The results of the convolution analysis can be easily plotted as a correction curve in the manner given by Jones[3] wherein β/B is plotted as a function of b/B. In order to simplify the terminology the notation of Jones will be used as far as possible. However, as it will prove necessary to refer to the different symbols of reference 6 occasionally, it will be helpful at the outset to compare the two notations:

Function	Jones	Reference 6
Pure diffraction breadth	β	w
Observed breadth for large crystallites (uncorrected for Kα doublet)	b_0	w_6
Same (corrected for Kα doublet)	b	w_5
Observed breadth for small crystallites (uncorrected for Kα doublet)	B_0	w_4
Same (corrected for Kα doublet)	B	w_3

It will be noticed that the contour breadths for small and large crystallites have not been distinguished in the notation of reference 6. This is because the analysis of reference 6 was of a general nature, applicable to contours of any breadth. Unless otherwise indicated the term *width*, or *breadth*, of a contour in the present paper denotes the width at half maximum intensity. Jones' analysis and correction curves apply to the integral breadth, which is defined by $\beta_z=(\int I_z dx)/(I_z)_{max}$.

The present analysis will deal mainly with the case of a relatively wide x-ray source (appropriate for the older Norelco spectrometer). For materials with crystallite size larger than 1000A the observed contour breadth b_0 will be due largely to the x-ray source and to a lesser extent to the receiving slit (unless an abnormally wide slit is employed). For materials smaller than 1000A the observed contour breadth B_0 will also be widened because of the width of the pure diffraction contour, β. If the crystallite size is very small the pure diffraction broadening becomes the dominant factor in determining the breadth of the observed contour. In addition to these effects, if the Kα doublet is not resolved, the observed contour will in all cases be widened as a result of the superposition of the overlapping contours produced by the α_1 and α_2 components of the radiation.

It will be assumed in the present paper that all observed contours have been corrected for Kα doublet broadening before the crystallite size correction curve is to be employed. The curve of Jones[3] for correcting for Kα doublet broadening is satisfactory for use with x-ray spectrometers having wide sources (of the order of 0.20° 2θ).

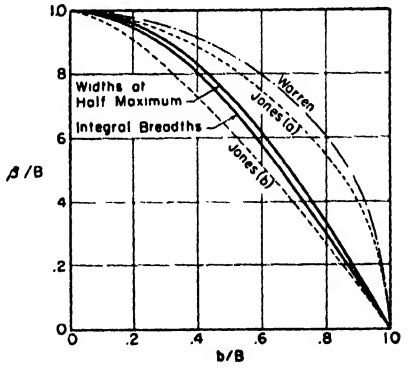


FIG. 1. X-ray spectrometer curves for correcting the breadths of diffraction contours for instrumental broadening. Also shown are the curves of Jones and Warren for Debye-Scherrer photographs.

Our procedure will now be as follows: (a) Using Eq. (1) compute the pure diffraction breadth, β, of the contour produced by a material of crystallite dimension D. (b) Using the convolution method of reference 6 compute the breadth B of the contour generated by the instrument without the added effect of the Kα doublet. (c) Similarly compute the breadth b of the contour generated for materials of large crystallite size (for which $\beta \to 0$). (d) Following the appropriate scheme of Jones, plot β/B as a function of b/B.

The success of this method of arriving at the correction curve will be contingent upon our correct choice of mathematical functions to describe the pure dif-

fraction contour and the x-ray source contour. A careful appraisal of these functions will be found in Parts III and IV of reference 6, so that it is necessary here only to repeat the final results, *viz.*, that a typical crystallite size distribution tends to generate a pure diffraction contour which is closely approximated by $1/(1+k^2\phi^2)$, whereas an x-ray source contour is more nearly described by $e^{-k^2\phi^2}$. Accordingly, the derivation of the correction curve which follows will be based on the premise that these simple mathematical functions accurately describe the pure diffraction and x-ray source contours.

The method of determining the curve of β/B as a function of b/B will be illustrated in detail by the calculation of one selected point, after which the complete data required for locating the entire curve will be tabulated. We shall select the case $\theta=30°$ and a crystallite size of $D=500$A. The spectrometer is assumed to be of the older design with an x-ray source width (w_I) of 0.200°, and a relatively narrow receiving slit width, $w_s=0.088°$, is employed. The radiation is CuKα, for which $\lambda=1.542$A.

Using the standard crystallite size equation with K set equal to unity, we find for the theoretical breadth of the pure diffraction contour

$$\beta=w=\frac{K\lambda}{D\cos\theta}=\frac{1.542\times57.3}{500\times.866}=0.204°.$$

The ratio of the width of the contour to that of the source is $w/w_I=0.204°/0.200°=1.02$. Referring to Fig. 4(I) of reference 6, we find for the broadening ratio $w_1/w=1.62$. Then the width of the resulting contour is

$$w_1=w\times(w_1/w)=0.204°\times1.62=0.330°.$$

This contour will be intermediate between $e^{-k^2\phi^2}$ and $1/(1+k^2\phi^2)$ in form because it is the convolution of two functions which are of these two pure forms and of approximately equal width.

We turn next to the action of the receiving slit in broadening this contour. The ratio of its width to that of the slit is $w_1/w_s=0.330°/0.088°=3.75$. Since we are neglecting the instrumental weight functions II, III, and IV, this ratio becomes the abscissa, w_4/w_s, of Fig. 4(V) of reference 6. The initial contour will be broadened by an amount which is approximately the mean of the values given by curves (a) and (b), hence $w_5/w_4=(1.030+1.036)/2=1.033$. This leads to a resulting contour width of

$$B=w_5=w_4\times(w_5/w_4)=0.330°\times1.033=0.341°.$$

For the reference material of large crystallite size the pure diffraction breadth β is zero. Hence, we start with the source contour and determine the broadening ratio due to the action of the receiving slit upon it. We have then

$$w_I/w_s=0.200°/0.088°=2.27,$$

TABLE II. Calculation of the crystallite sizes of four materials from spectrometer measurements after determining β (A) by Stokes' Fourier method and (B) from the correction curve of Fig. 1 (widths at half maximum).

	CaF_2	MgO	Cristobalite	Basic calcium phosphate
Observed breadths				
B_0	0.500°	0.700°	0.470°	0.468°
b_0 (quartz powder)	0.330°	0.330°	0.320°	0.330°
Breadths after correcting for $K\alpha_1\alpha_2$				
B	0.460°	0.674°	0.462°	0.456°
b	0.276°	0.284°	0.307°	0.314°
Correction ratios				
b/B	0.600	0.421	0.665	0.689
β/B (Fig. 1)	0.620	0.810	0.533	0.500
Pure diffraction breadths				
β (Stokes' method)	0.335°	0.530°	0.260°	0.215°
β (Fig. 1)	0.285°	0.546°	0.246°	0.228°
Calculated crystallite sizes				
D(A.) (Stokes' method)	290	179	346	422
D(A.) (Fig. 1)	342	174	366	398

and from Fig. 4(V), curve (a), of reference (6), $w_5/w_4=1.05$, and consequently

$$b=w_5=w_4\times(w_5/w_4)=0.200°\times1.05=0.210°.$$

From these calculated values of β, b, and B we obtain

$$b/B=0.616 \quad \text{and} \quad \beta/B=0.598.$$

By proceeding in a similar manner for a number of crystallite sizes ranging from 50 to 1000A the remaining data of Table I of the present paper were obtained. When plotted with β/B as a function of b/B, the solid curve marked "Widths at Half Maximum" of Fig. 1 is obtained. Although deduced specifically for the case of $\theta=30°$ and a receiving slit width of 0.088°, this curve is found to be perfectly general for all values of θ in the range 0–45° and for receiving slit widths ranging from 0 to about 0.20°. When the slit width exceeds the angular width of the source by an appreciable amount, the correction curve shifts to the right. The most satisfactory policy in this regard is to always employ the narrowest receiving slit that is compatible with the diffracted energy available.

By a parallel set of calculations it can be shown, or by direct comparison of the geometrical conditions it can be inferred, that this same correction curve applies, within limits, to spectrometers with narrower sources. However, when the source width becomes much smaller than 0.10°, the several minor broadening factors become relatively significant, so that the premises upon which the present curve is based are no longer correct.

III. EXPERIMENTAL SUBSTANTIATION OF THE CORRECTION CURVE

In Part III of reference 6 the pure diffraction contours of six miscellaneous materials were investigated, five of them by the Fourier transform method of Stokes.[5] Four of these materials made excellent subjects

TABLE III. Comparison of the crystallite sizes of four materials as deduced by different correction curves. The figures are calculated crystallite sizes in A.

Method of correction	CaF_2	MgO	Cristobalite	Basic calcium phosphate
Stokes (Fourier)	290	179	346	422
Curve of Fig. 1, widths at half maximum	342	174	366	398
Jones' curve (a)	285	162	284	301
Jones' curve (b)	414	201	448	491
Warren's curve $[\beta/B=(1-b^2/B^2)^{\frac{1}{2}}]$	265	156	260	274

for a test of the spectrometer correction curve of Fig. 1. The widths of the pure contours deduced by the Fourier method can be considered to be the best obtainable and hence to also lead to the most reliable values of the crystallite sizes. In Table II the crystallite sizes are calculated using the curve of Fig. 1 marked "Widths at Half Maximum," and the resulting values are compared with the presumably best values from the Stokes procedure.

In every case an appropriate quartz maximum was measured in order to obtain the required value of b_0 for a material of large crystallite size. The contours were measured with a Norelco Geiger-Counter Spectrometer, Type No. 12021 (with wide x-ray source), the spectrometer arm being set manually at appropriate intervals and the counting rates being recorded directly. In every instance except that of basic calcium phosphate it was possible to reduce the probable statistical counting error in the peak height above background to 1 percent or less. It is seen that in three out of four cases the agreement is highly satisfactory, and, furthermore, that in the single case of less satisfactory agreement (CaF_2) a greater deviation from the presumably best value (by Stokes' method) is to be expected because the pure diffraction contour is intermediate in nature between the forms $1/(1+k^2\phi^2)$ and $e^{-k^2\phi^2}$ rather than resembling more closely the first function (refer to Table I of reference 6).

Table III compares the crystallite sizes of these same materials as calculated using various correction curves with the values arrived at by Stokes' Fourier method. For the sake of ready comparison these correction curves are also plotted in Fig. 1. An examination of Table III shows that on the average the present curve gives much better agreement with the Fourier results than any of the other curves. Only in the case of the somewhat anomalous contour of CaF_2 would somewhat better results have been obtained by using other curves, e.g., by employing either Warren's curve or Jones' curve (a). It is interesting to note from Fig. 1 that the new spectrometer curve agrees more nearly with Jones' curve (b) at large values of b/B (large crystallite sizes), while it approaches both Jones' curve (a) and Warren's curve more closely at small values of b/B (small crystallite sizes). However, as is well-known, the choice of correction curve is unimportant for small crystallite sizes.

IV. DISCUSSION OF THE CORRECTION CURVE

In Part II it was emphasized that the derivation of the present correction curve is based on the premise that the functions $1/(1+k^2\phi^2)$ and $e^{-k^2\phi^2}$ accurately describe the pure diffraction contour and x-ray source contour respectively. These formulas should be regarded realistically as being only limiting functions which are approached only more or less imperfectly in any experimental situation. Consequently the most that should be claimed for the present correction curve is that the probability of its yielding an accurate value of β is good and also much better than that of curves based on other pairs of simple mathematical functions.

In particular, it is suggested that each investigator test the focal spot of the particular x-ray tube employed for agreement with the function $e^{-k^2\phi^2}$. A simple but sufficiently accurate method for accomplishing this is described in Part IV of reference 6. If reasonably good agreement is found he may then use the present correction curve with confidence.

For a spectrometer with a wide source the correction curve of Fig. 1 will have rather general reliability provided that the following experimental requirements are satisfied: (a) The linear absorption coefficient of the powder (including solid and interstices) should be approximately 40 or larger. (b) The diffraction angle, 2θ, should be greater than 20°. (c) The receiving slit width should not exceed the width at half maximum intensity of the x-ray source contour (both expressed in terms of 2θ units). (d) By proper adjustment of the x-ray slit the horizontal and vertical divergence of the x-ray beam should preferably be limited to moderate values, say, 1° and 2° respectively. Under these circumstances, if we disregard the usual remaining uncertainty in K of Eq. (1), satisfactory measurements can be made of crystallite sizes as large as 500A provided that B_0 and b_0 can be measured with an accuracy of ±0.01° or better for the larger sizes. For the new spectrometers with narrower x-ray sources the present curve is fairly reliable for crystallite size measurements below 500A. However, although instruments of this design are theoretically capable of application in the size range 500–1000A, the novel geometrical factors involved are such as to require a special analysis in order that a correction curve may be established for work in this size range.

If it is desired to use integral contour breadths, defined by $\beta_x=(\int I_x dx)/(I_x)_{max}$, rather than widths at half maximum intensity, the curve marked "Integral Breadths" in Fig. 1 should be employed. This curve was derived by substituting the functions $1/(1+k_1^2x^2)$ and $\exp(-k_2^2x^2)$ respectively for $F(kx)$ and $f(x)$ in Jones' Eqs. (4) and (5).[8] The integral breadth is more convenient for mathematical analysis but less readily employed in the experimental measurement of crystallite size.

When measurements of the highest reliability are required, it is recommended that the Fourier transform method of Stokes be employed for determining β. Once the equipment for the rapid summation of Fourier series has been constructed and the technique for performing the operations has been mastered, the time consumption per analysis is not excessive (possibly one or two hours). The Fourier method has completely

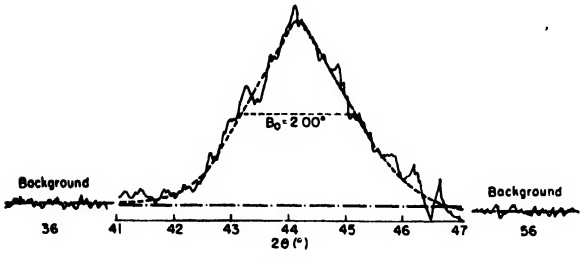


FIG. 2. Determination of the crystallite size of a nickel catalyst from an automatic strip chart record of the (111) maximum. CuKα-radiation; scanning speed 1° per minute; background measured at $2\theta=36°$ and 56°. Neglecting b_0, $D_{111}=48$ A. Using $b_0=0.33°$ for the (220) reflection of NaCl powder, $D_{111}=49$ A.

general applicability regardless of the particular forms of the pure diffraction contour and the various instrumental broadening functions. Furthermore Kα doublet broadening is automatically allowed for.

V. PRACTICAL SUGGESTIONS FOR CRYSTALLITE SIZE DETERMINATION WITH THE X-RAY SPECTROMETER

The following discussion applies to the older spectrometer design incorporating a wide x-ray source and a working range of $2\theta=0$ to 90°. Three crystallite size ranges will be given special consideration, each amenable to a somewhat different experimental approach.

For all except the smallest crystallite sizes it is necessary to measure the contour widths of both the material being studied (B_0) and a reference material which gives no crystallite size broadening (b_0). In order that the geometrical broadening effects may be as nearly identical as possible, the two peaks should be situated at approximately the same diffraction angle. When the linear absorption coefficients for x-rays of the "unknown" and reference powders are both high (say, $\mu=40$ or more when the interstices of the powder are included in the calculation) or when the absorption coefficients are known to be very nearly equal, good results can be obtained when the two materials are mounted and analyzed separately. However, when the coefficients are small, and in particular when they are known to differ considerably, the two materials should be intimately mixed and the measurements made on a single sample. The fact that the calculated absorption coefficients of the two solid materials are nearly equal is in itself no guarantee that the powders will absorb equally because they may not pack to the same density. When this is evidently true, or when there is any doubt in this regard, the materials should always be mixed. Fine quartz powder manifests a fairly typical packing behavior and is found to assume a state in which solid and interstices constitute about 40 and 60 percent respectively of the total sample volume. Hence, the linear absorption coefficient of the powder is only 0.40 as large as that of the solid material [for CuKα radiation μ (quartz powder) $=0.40\times93=37$].

The above suggestions presuppose the use of a sample of sufficient thickness to give maximum absorption (see remarks in reference 6, Part VII). If sufficient intensity can be obtained with the use of a very thin sample, such as a film of powder mounted on a suitable plane substrate, the differences in absorption become much less important and it is usually unnecessary to mix the powders.

Case 1. Size Range < 150A

In this range fairly good results can be had by preparing automatic strip chart records at a sufficiently slow scanning speed to yield a contour of easily measurable width. The background is also measured at selected points on either side of the maximum, after which the best straight horizontal line is drawn through the statistical fluctuations of each background record and a smooth curve is drawn to delineate the peak contour. This method is not successful unless the peaks in question are relatively strong. In favorable cases of this sort β can be determined with a probable error of 5 to 10 percent. Figure 2 illustrates the application of this method to a nickel catalyst of fine crystallite size.

If the peak contours are counted and plotted manually and if sufficient counts are taken to permit the

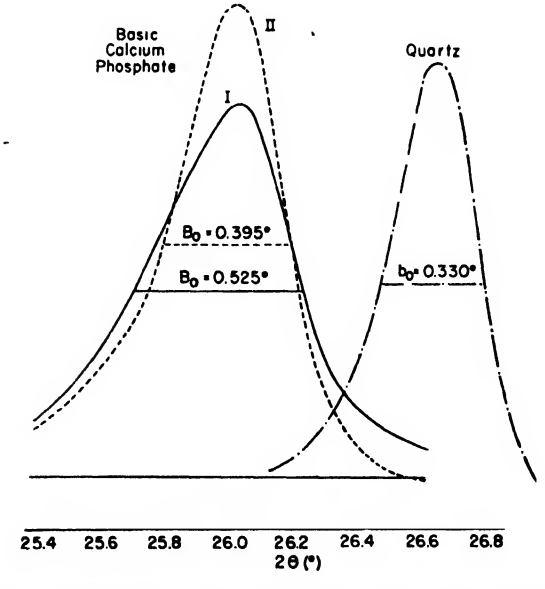


FIG. 3. Diffraction contours of basic calcium phosphate precipitate before (I) and after (II) digestion, and reference contour of quartz powder. The contours were plotted from manually recorded counting rate data. Crystallite sizes (after correcting for Kα doublet broadening): (I) 290 A, (II) approximately 800 A.

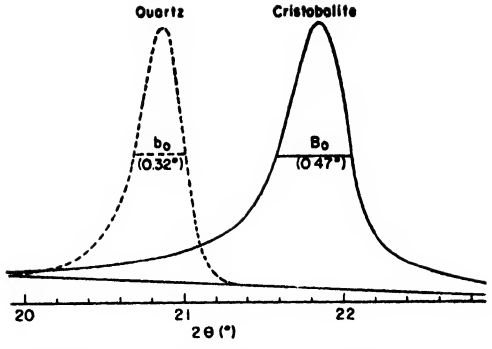


FIG. 4. Diffraction contours used in determining the crystallite size of cristobalite prepared by calcining diatomaceous earth. The contours were plotted from manually recorded counting rate data. Mean crystallite size (after correcting b_0 and B_0 for Kα doublet broadening), 366 A.

determination of the peak height above background with a probable error from counting statistics of 1 percent or less (5000 counts per point for cases with negligibly small backgrounds), the probable error in β can be reduced to 3 percent or less. In this size range it is unnecessary to correct the observed line breadths for Kα doublet broadening. Warren[2] has pointed out that the measured breadth of the "unknown" line, B_0, can be substituted for β directly in the crystallite size formula when $b_0 < B_0/5$, because in such instances the broadening due to the instrumental effects becomes insignificant in comparison with the pure diffraction broadening. Thus, for the nickel catalyst of Fig. 2 the calculated crystallite size was 49A when b_0 was taken into account and 48A when not.

Case 2. Size Range 150–300A

Automatic recording is not satisfactory in this range except for very rough measurements. If the heights of the peaks above background are determined with an accuracy approaching 1 percent (as limited by counting statistics), it is possible to determine β with a probable error of 5 percent. This and succeeding estimates of precision and error take into account a reasonable degree of uncertainty in the position of the spectrometer correction curve (Fig. 1). In this size range it is unnecessary to correct for Kα doublet broadening for $2\theta < 60°$. For higher angles the error involved is sufficient to increase appreciably the 5 percent probable error quoted above.

Case 3. Size Range 300–500A

In this range it is necessary to adopt the best experimental techniques in order to secure good results. A probable error in β of considerably less than 10 percent can be attained by keeping the probable error in counting statistics at points near the tops of the peaks smaller than 1 percent and by correcting for Kα doublet broadening *in all cases*. In addition, for the most accurate work it is recommended that the recorded counting rates be corrected for the nonlinearity of response of the Geiger tube.[7]

Of course, care must also be exercised in defining accurately the sides of the contours at the levels at which the breadths B_0 and b_0 are to be measured. This means measuring the counting rates at two or more points in the immediate vicinities. An analysis of the counting errors involved has shown that it is satisfactory to use a constant time interval for counting all the points defining a given contour, provided only that this time interval furnishes a sufficient number of counts to determine the height of the peak above background with high precision.

The background should preferably be counted with a longer time interval because of the normally much smaller intensities involved, and it is suggested that this interval be made twice as long as that employed in counting the peak contour. When the background is abnormally high and constitutes an appreciable portion of the over-all counting rate measured at the top of the peak, it is necessary to totalize much more than 5000 counts at the top of the peak in order to reduce the probable error in the *net* peak height to less than 1 percent.[8] In any event care should also be taken that the background be measured at points sufficiently far removed from the peak to avoid contributions from the base of the peak itself. For sharp peaks an angular displacement from the peak center of 2.0° is usually satisfactory, but for the very wide maxima generated by a distribution of crystallite sizes in the range below 100A it is necessary to increase this to 5° or even more.

Figures 3 and 4 illustrate crystallite size measurements in the ranges of Cases 2 and 3. In each figure the required reference contour of a material of large crystallite size (fine quartz powder) is also shown. Actually each contour pictured was obtained using a separate sample, but the several contours have been drawn together to facilitate comparison. The growth of the basic calcium phosphate crystallites during digestion is strikingly reflected in the alteration of the diffraction contour shown in Fig. 3. The accuracy of the measurement following digestion is obviously not good (probably not better than 800±150A) because the size involved is outside the satisfactory working range of the spectrometer. Figure 4 portrays the b_0 and B_0 contours which were analyzed in determining the crystallite size of cristobalite, a form of crystalline silica, prepared by calcining diatomaceous earth.

ACKNOWLEDGMENT

The authors are indebted to Elizabeth Kummer and Douglas Pitman for assistance in obtaining the experimental data and performing calculations.

[7] Alexander, Kummer, and Klug, J. App. Phys. 20, 735 (1949).
[8] Tracerlog No. 2, p. 5, March 1947.

Radiofrequency Mass Spectrometer

WILLARD H. BENNETT
National Bureau of Standards, Washington, D. C.
(Received August 29, 1949)

A mass spectrometer tube using velocity selection is described. The tube has parallel plane grids made with knitted wire nets with a large percentage of open area. The grids are arranged in groups of three, and radiofrequency alternating potential is applied to the middle grid of each group. Stopping potentials are used to turn back all ions except those with the selected mass. The spectrometer has sufficient resolution for ordinary gas analysis requirements and is simpler, more compact, and more rugged than magnetic beam deflection devices.

SEVERAL proposals have appeared in the past for making a mass spectrometer using velocity selection instead of a magnetic deflection of a focused beam of ions.

Smythe and Mattauch[1] used a succession of pairs of condenser plates along a tube, with which radiofrequency alternating fields could be applied to the ions in a collimated beam at right angles to the principal direction of movement of the ions in the beam. The ions in order to arrive at the receiver had to have the velocity necessary to place them opposite each successive pair of condenser plates at successively opposite phase. Because harmonic masses would also be recorded in such an arrangement, Smythe and Mattauch[1] also bent the beam through an arc with an electric field between a curved pair of electrodes in order to eliminate harmonics. From the available publications, this method does not appear to have been successful.

Stephens[2] proposed, and Cameron and Eggers[3] built and tried a pulsed ion source in combination with an oscilloscopic display of the distribution in time of the arrival of ions at a receiver. The resolution obtained, using this attractively simple idea, is disappointingly limited. This limitation is probably inherent in any method which has to use such high speed recording of such small pulses of current.

Weisz[4] has suggested, but apparently not tried, an adaptation of a linear accelerator in which alternately phased radiofrequency alternating potential is applied to a succession of equally spaced coaxial cylindrical electrodes. He proposes to operate the instrument in a manner such that the ions to be selected arrive at the gaps between electrodes at zero phase and that they be selected on the basis that they have neither gained nor lost energy because of the alternating fields between the successive cylindrical electrodes. He fails to make clear how to accomplish such a selection.

Bauer[5] has suggested a method which requires a well-focused beam to travel between parallel plates between which an alternating radiofrequency potential is applied. This method also appears not to have been tried.

Proposals have also appeared[6] for time-of-flight methods which use uniform magnetic fields in obtaining well-focused mass components, and as such have most of the disadvantages of the more familiar magnetic beam deflection methods and can use only as intense an ion beam as can be made to pass through a narrow slit.

The present development grew out of some observations made in 1946 on a pentode with coaxial cylindrical electrodes. A radiofrequency a.c. potential was applied to the intermediate one of the three grids while negative ions were produced at the hot cathode in the central portion of the tube. Although this tube was built for the purpose of separating kinds of negative ions differing in mass by a factor of two or more, it was found that a much better separation was effected than had been anticipated. For this reason, experiments were initiated to explore the possibilities of this kind of velocity selector as a mass spectrometer.

SINGLE-STAGE TUBE

The first tube built using parallel plane grids was a single-stage tube, as shown in Fig. 1, and illustrated schematically in Fig. 2. Electrons from the filament, *a*, are accelerated through a d.c. potential to the first grid, *b*. Positive ions produced by collisions between these electrons and gas molecules in the tube between grids *b* and *c* are accelerated from *b* toward *c* with a potential difference, *V*, between *b* and *c*. The grids *c*, *d*, and *e* are equally spaced at a distance *s*. A radiofrequency field of angular frequency ω is applied to the middle grid *d* giving an electric field between *c* and *d*

$$E_{cd}=E\sin(\omega t+\theta)$$

and between *d* and *e*

$$E_{de}=-E\sin(\omega t+\theta);$$

where t is zero at the instant an ion crosses the plane of grid *c* and $E\sin\theta$ is the value of the field E_{cd} at that instant. We assume a uniform flow of ions up to grid *c* so that there will be equal numbers of ions with each

[1] W. R. Smythe and J. Mattauch, Phys. Rev. 40, 429 (1932).
[2] W. E. Stephens, Phys. Rev. 69, 691(A) (1946).
[3] A. E. Cameron and D. F. Eggers, Jr., Rev. Sci. Inst. 19, 605 (1948).
[4] P. B. Weisz, Phys. Rev. 70, 91(L) (1946).
[5] S. H. Bauer, J. Phys. Chem. 39, 959 (1935).
[6] S. A. Goudsmit, Phys. Rev. 74, 622 (1948). J. A. Hipple and H. A. Thomas, Phys. Rev. 75, 1616 (1949).

possible value of θ. Some of these will receive energy from the fields as they pass the grids while others will lose energy to the fields. The collecting plate f has a d.c. stopping potential which turns back all ions except those which have acquired nearly the maximum possible energy from the fields and which originated near the grid b.

An exact solution for the motion of a charged particle in the sinusoidal field is extremely complex and has not been obtained. For purposes of this discussion, a small disturbance approximation will suffice. Experiment has shown that this approximation holds quite well for finite amplitudes. It is assumed that the change in velocity of an ion as it moves through the grids is small compared to its initial velocity so that the effects

FIG. 1.

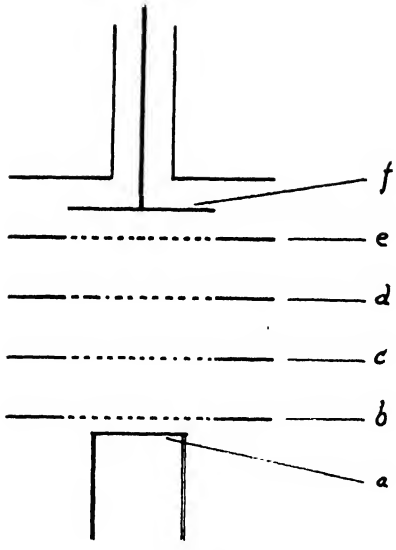


FIG. 2.

of the applied fields on the transit time can be neglected. Space-charge effects are also neglected with negligible error for this system.

The force acting on mass m between grids c and d is

$$F_{cd}=m\ddot{x}=Ee\sin(\omega t+\theta)$$

and between d and e is

$$F_{de}=m\ddot{x}=-Ee\sin(\omega t+\theta),$$

where time t is measured from the time the ion passes the grid c and θ is the phase angle of the alternating potential on d at the instant the ion passes c. Taking for a first approximation, the case when the amplitude of the alternating potential is small compared with the d.c. potential, V, used in accelerating the ion from the source to the grid, c, the variations in the velocity of the ion due to the a.c. fields will be small compared with the average velocity of the ion between grids c and e. The energy acquired from the alternating field is approximately:

$$\Delta W=\Delta(mv^2/2)=v\cdot\Delta(mv)=v\int F dt.$$

The transit time for the ion to travel from one grid to the next is s/v where s is the distance between grids. Then

$$\Delta W=v\left\{\int_0^{s/v}Ee\sin(\omega t+\theta)dt+\int_{s/v}^{2s/v}-Ee\sin(\omega t+\theta)dt\right\}$$

$$\Delta W=\frac{Eev}{\omega}\left[\cos\theta-2\cos\left(\frac{s\omega}{v}+\theta\right)+\cos\left(\frac{2s\omega}{v}+\theta\right)\right].$$

This is a maximum with respect to variations in θ when

$$(s\omega)/v+\theta=\pi=180°,$$

showing that the ions which pass grid d just as the field reverses obtain maximum incremental energy. The expression for ΔW is a maximum with respect to variations in ω when

$$\theta = 46° \text{ (more precisely } 46°26')$$

so the transit angle between grids is

$$(s\omega)/v = 134°.$$

The velocity of the ion receiving maximum incremental energy is obtained from

$$eV = \tfrac{1}{2} M m_0 v^2,$$

where V is the potential difference between the ion source, b, and the grid, c; M is the atomic mass number of the ion; and m_0 is the mass of an ion of unit mass number. Eliminating v between the last two equations gives

$$M = \frac{0.266.10^{12} V}{s^2 f^2},$$

where V is in volts, s is in centimeters, and f is in cycles per second.

Shown in Fig. 3 is the energy, ΔW, acquired by an ion passing the grid d, at 180° phase as a function of the number of cycles, N, executed by the alternating field while the ion travels from grid c to grid e. The principal maximum of this curve is at $N = 0.74$ cycle because the transit angle from the first to the third grid for maximum ΔW is

$$(2s\omega)/v = 267° = 0.74 \text{ cycle}.$$

If the stopping potential applied in the collecting system corresponds to the energy Z, at the horizontal dashed line, the ions can be collected only at frequencies corresponding to positions in the graph between A and B. Ions passing the first grid at phase angles different from the optimum 46° angle acquire a lower energy and can reach the collecting electrode within a narrower range than the range between A and B. The ions collected are those at all phases of entry which can collect enough ΔW to exceed Z so the shape of the line observed for ions of a particular mass is sharper than that portion of the curve above Z.

An example of the line form observed is shown in Fig. 4. This is a plot of the experimentally observed ion currents for mercury positive ions. A smaller stopping potential is used for the upper curve than for the lower curve.

It was found by experiment that the above approximate treatment was adequate if the d.c. potential difference between the source and the first grid was more than 10 times the amplitude of the r-f alternating potential, but that irregular curves for each mass component resulted at larger relative a.c. potentials. However, symmetric mass line forms resulted for a.c. potentials up to about half the d.c. potential on grid d, to which the a.c. potential is added, is reduced by just the amount necessary to reduce the speed of an ion receiving maximum incremental energy, back to the same speed as that with which the ion passed grid c, and if the d.c. potential on grid e were also similarly reduced below that of grid d.

A good example of this can be given in the case where the tube is used for resolving the negative atomic ions produced at the surface of a hot cathode. In this case the potentials applied were:

Ion source, a,	0.0 volt, d.c.
Grid, b,	+10.0 volts d.c.
Grid, c,	+ 4.0 volts d.c. plus 6 volts (r.m.s.) a.c.
Grid, d,	− 2.0 volts d.c.
Grid, e,	−14.1 volts d.c.
Collector, f	+20.0 volts d.c.

It will be noted that in this particular test, the stopping potential was applied to the fourth grid, e, and not to the collector. A simple integration shows that if the grids were ideal grids (very small holes compared with the inter-grid distance), the maximum incremental energy an ion could acquire from the r-f fields would be about the same as would be acquired from a d.c. potential twice the r.m.s. value of the r-f potential on the intermediate grid. Thus the use of an r-f potential of six volts requires stepping back the d.c. potential on each successive grid by about six volts in order to hold approximately constant the velocity of those ions receiving nearly maximum incremental energy. The value of the stopping potential, −14.1 volts, exceeds the value of a little less than −12 expected because the more positive potential of the adjacent electrodes reduces the magnitude of the potential of the middles of the holes in the stopping grid to less than that of the potential applied to that grid.

By appropriately selecting the d.c. potentials, this spectrometer tube serves equally well for separating positive or negative ions. Although no magnetic field is needed with this tube in separating positive ions, it is helpful to use a small magnetic field perpendicular to

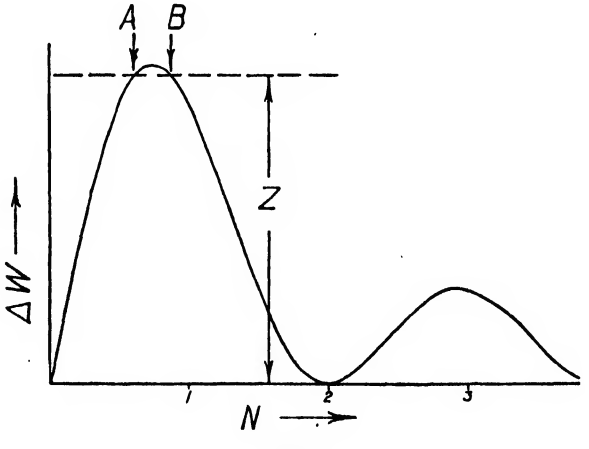


FIG. 3.

the axis of the tube when separating negative ions in order to confine the electrons from the source to the vicinity of the source. This prevents the formation of positive ions throughout the tube by the relatively large electron currents which would otherwise reach the vicinity of the stopping grid.

TWO-STAGE TUBE

The method was extended to a two-stage tube shown schematically in Fig. 5. In this case, the grids *b*, *c*, and *d* serve as the first stage while grids *e*, *f*, and *g* function as the second stage. The same radiofrequency alternating potential is applied to grids *c* and *f* in addition to the two unequal d.c. potentials whose difference provides the compensation described previously in connection with the single-stage tube. As before, the potential *V* accelerates the ions from the source *a* to the first grid *b*, and the d.c. potentials of grids *c*, *d*, *f*, and *g* are similarly adjusted. The grids *d* and *e* are at the same potential so that the region between them is an equipotential drift space. The stopping grid *h* and the collecting plate *p* function as before.

The distance between grids *d* and *e* is such that an ion receiving maximum incremental energy from the first stage will reach the initial grid of the second stage, *e*, at the right phase of the r-f alternating potential to acquire a maximum incremental energy from the second stage also. In order to accomplish this, the r-f potential must execute exactly an integral number of cycles while the ion travels from stage to stage, or more exactly, from *c* to *f*, and so the distance from *d* to *e* must be

$$(2.70n-2)s,$$

where *n* is an integer, and *s* is the distance between the grids in each stage as before. The first two-stage tube was built with $n=6$. This will be referred to as a six-cycle two-stage tube.

The incremental energy which can be acquired by an ion from both stages of a two-stage tube is illustrated in Fig. 6. The envelope of the curve is the same kind of a curve as plotted in Fig. 3 but with twice the ordinate, of course.

In Fig. 7 is shown the current of negative iodine ions from a hot tungsten filament in iodine vapor at 10^{-4}

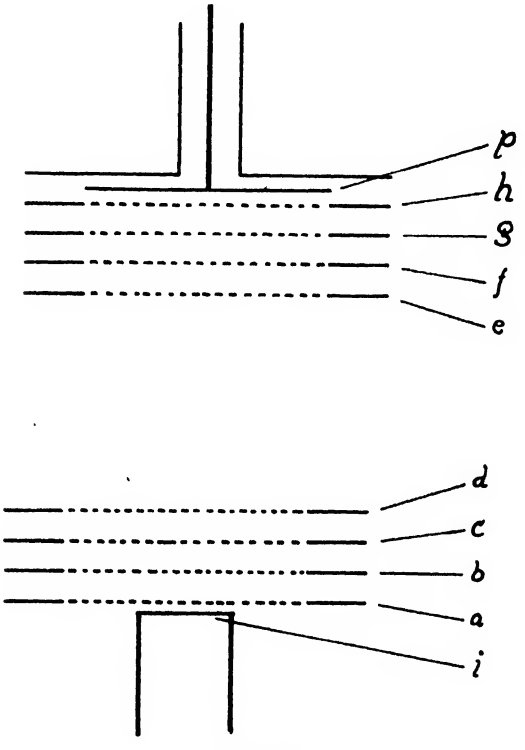


FIG. 5.

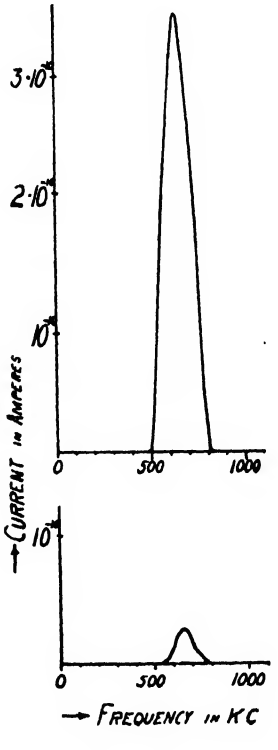


FIG. 4.

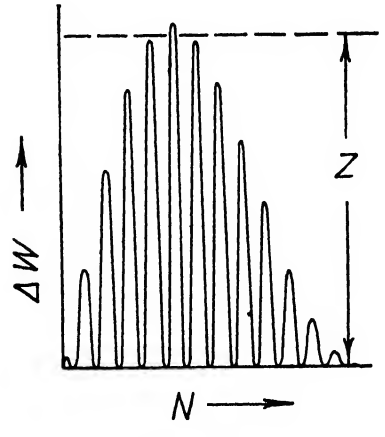


FIG. 6.

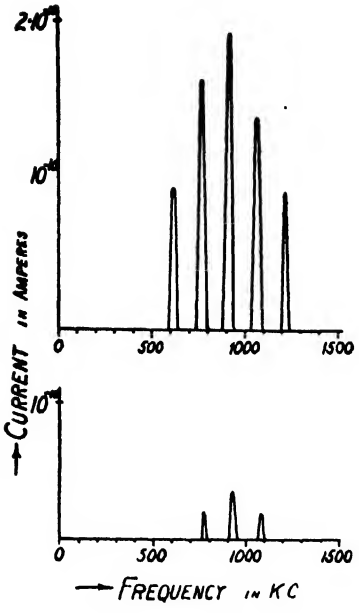


FIG. 7.

mm pressure. The lower curve shows the currents observed using a larger stopping potential, Z, than that used in obtaining the upper curve. It is seen by comparing these curves with those in Fig. 4 that the selection is much sharper than that obtained with a single-stage tube. High peaks occur for frequencies corresponding to other numbers of cycles between stages than the number 6 for the principal peak, and much of the sensitivity of the tube is lost in increasing the stopping potential enough to cut out the adjacent peaks.

THREE-STAGE TUBE

In order to retain much more sensitivity and, at the same time, produce only a single peak for each mass, a third stage is introduced. Such a tube is shown in Fig. 8. The three stages must be spaced so that the r-f alternating potential on the middle grid of each stage executes exactly integral numbers of cycles while the ion acquiring maximum incremental energy travels from stage to stage, i.e., from e to h, and from h to k. These integral numbers between the two successive pairs of stages are so selected as to avoid any appreciable overlap of the harmonic peaks of the two pairs of stages considered as two-stage combinations. This is illustrated in Fig. 9. When the integral numbers are so selected, much more of the principal peak can be left above the blocking potential, Z, than with either the single-stage or two-stage tubes and a corresponding increase in sensitivity is effected.

A 9–7-cycle three-stage tube was built, shown in Fig. 8, and illustrated schematically in Fig. 10. This tube was provided with two different kinds of ion source. One made use of a pencil of electrons, a', which passed at right angles to the axis of the tube. Ions produced along this electron beam were made to drift toward the grid c with a small electric field between b and c. The other source of ions which could alternatively be used made use of electrons from filament a accelerated toward grid b. Ions produced between b and c were similarly made to drift toward c with a small field between b and c.

A d.c. potential, V, accelerates the positive ions emerging through c toward and through d. The grids d, e, and f act as the first stage; g, h, and i the second stage; and j, k, and l the third stage. The only grids receiving the radiofrequency a.c. potential are e, h, and k. The stopping potential difference is applied between the collecting plate p and the grid c. The grid m was

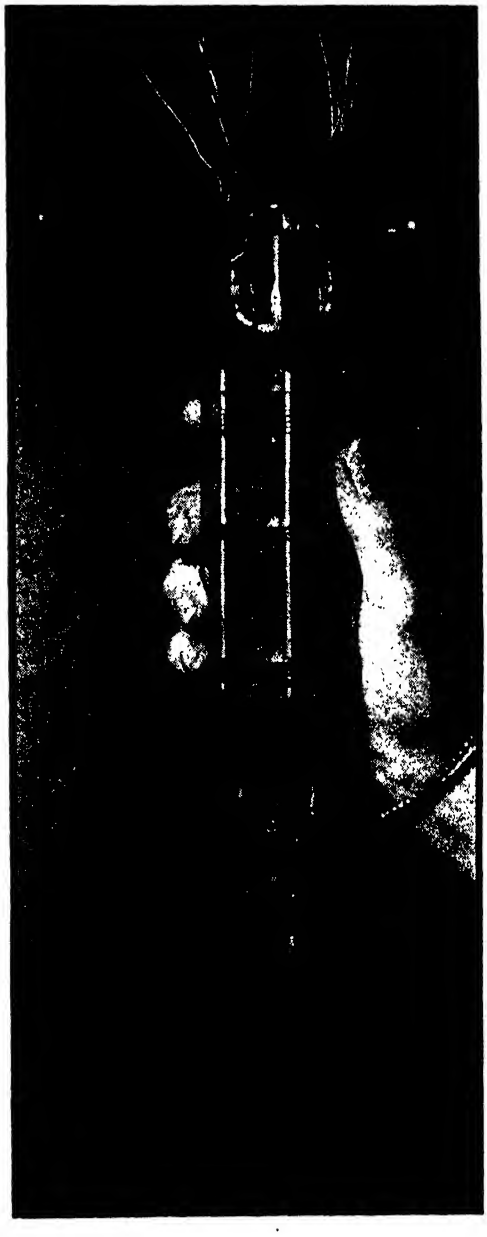


FIG. 8.

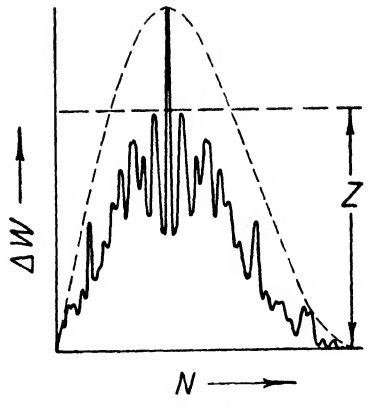


FIG. 9.

held at the same potential as the collector, and grid *n* was held at a more negative potential than any other electrode in the tube in order to turn back all electrons produced in the tube between grids *d* and *n*.

The grids used in this work were made with 0.0005-in. tungsten wire knitted with about 15 loops per centimeter. One of these grids is pictured in Fig. 11. The wire is mounted on the frame in a stretched condition so that the grid does not sag during or after baking. This is quite important in order that the grids remain accurately positioned with respect to each other. The open area of these grids is greater than 95 percent and even though 13 are used in a tube, more than 50 percent of the beam can pass through to the collector. The most open of the woven wire screens available with openings this small have open areas less than 65 percent, and 13 of these in series would pass less than one-half percent and would be quite impractical. Mr. William B. Haliday of this Bureau has assisted considerably in the development of the techniques for making these grids.

In this first of the three-stage tubes built, the grids in the stages were spaced at 2 mm which is of the same order of magnitude as the size of the holes in the grids. For this reason, the grids were far from ideal grids.

Figure 12 shows one of the observations of the spectrum of positive ions of mercury. In this observation, harmonic peaks occur at frequencies above and below the principal peak by about one-eighth the frequency of the principal peak. These harmonic peaks are evidently due to a partial overlap of the first harmonic peaks of the two two-stage combinations whose harmonics occur at frequencies, respectively, at one-ninth and one-seventh of the principal frequency, as illustrated in connection with Fig. 7.

Increasing the stopping potential and correspondingly reducing the intensity of the observed line eliminates these harmonic peaks as shown in Fig. 13. In this figure, the solid vertical lines under the curve are drawn proportional to the relative abundance of the isotopes. The width of this observed line at half the maximum height is 3½ percent while the width, due to the spread in isotope distribution, should be at least 1½ percent. However, the line flares out at the base much more than the lines did in the two-stage cases. This is believed to be due to the fact that this tube used grid

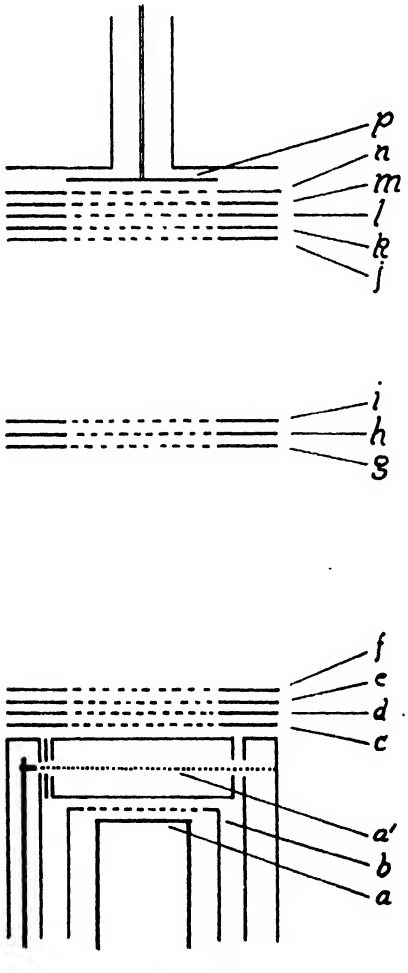


FIG. 10.

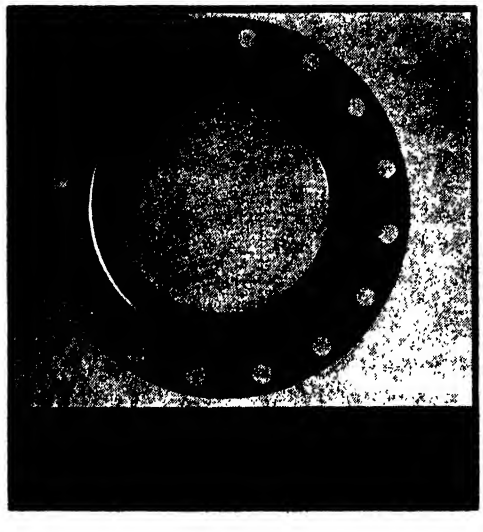


FIG. 11.

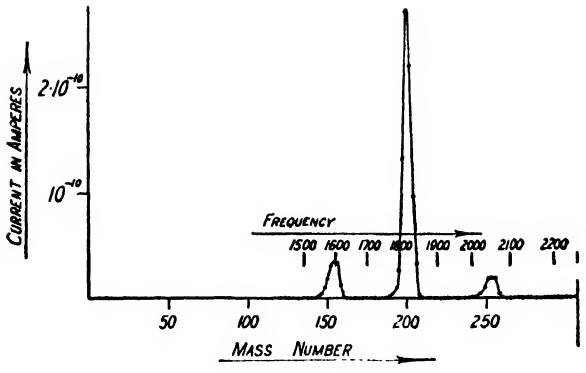


FIG. 12.

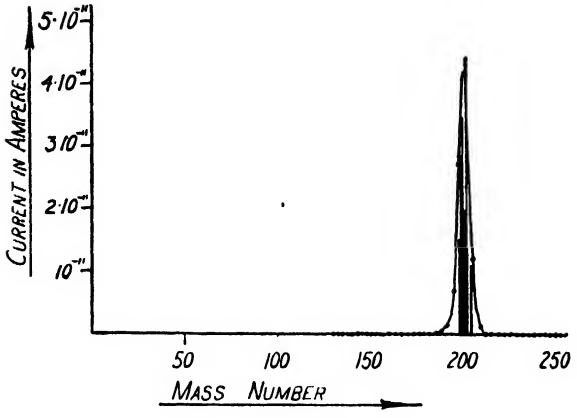


FIG. 13.

distances which were too small compared with the size of the holes in the grids.

From the shape of the line shown in Fig. 13, it is seen that this spectrometer should have sufficient resolution for those gas analyses where the masses to be completely separated differ by more than six percent, as is the case in many ordinary gas analyses of the lighter substances.

It should be noted that after constructing this spectrometer tube with proper care given to the accuracy of the grid spacings, and the accuracy of measurement of the potential, V, the formula

$$M = \frac{0.266.10^{12}V}{s^2f^2}$$

accurately gives the frequencies at which the various mass components appear. The instrument, when built with due attention to grid spacing, does not have to be calibrated with known gases and, in this sense, this is an absolute instrument.

This spectrometer tube has been found to operate well at pressures from 10^{-6} to 10^{-4} mm. It should be built in a form which can be baked because, otherwise, the gases usually found in unbaked tubes can give strong background mass components in the observations. On the other hand, masses can be resolved at pressures up to values of the pressure where the mean free path of the ions is about equal to the length of the tube.

It is important to note that no slits are required. All the ions produced in the ion source are accelerated through the essentially open structure of the tube toward the stopping electrode. The current at the top of the principal peak in Fig. 12 corresponds to about five percent of all the ions produced in the ion source.

Experiments are continuing to determine whether the harmonic peaks shown in Fig. 12 can better be eliminated by some other selection of integer combination for the number of cycles between stages or by increasing the ratio of the grid distances to the hole-sizes in the grids. A number of simplifications of the tube and equipment are now being tested to determine which can be introduced without losing too much resolution for analytical applications.

Mr. Harry Diamond and Dr. Robert D. Huntoon, of this Bureau, have contributed much in encouragement in this development.

Strain Induced Grain Boundary Migration in High Purity Aluminum

PAUL A. BECK AND PHILIP R. SPERRY*
Department of Metallurgy, University of Notre Dame, Notre Dame, Indiana
(Received August 8, 1949)

A method has been developed for the study of grain boundary migration in high purity aluminum, which allows positive determination of the *direction* of grain boundary movements. By means of this method it was confirmed that the migration of grain boundaries resulting from surface energy takes place in the direction toward the centers of curvature of the boundaries. On the other hand, strain induced grain boundary migration, such as occurs in recrystallization, proceeds in a direction away from the centers of curvature of the moving boundaries. It was also found that certain strained grains serve as nuclei for strain-free grains of the same orientation which grow at the expense of neighboring strained grains. Most of the strain-free grains formed during the annealing of high purity aluminum cold rolled up to about 40 percent are produced by this mechanism of strain-induced boundary migration, without the formation of new nuclei.

RECENT investigations of Harker and Parker,[1] C. S. Smith,[2] and J. E. Burke[3] revealed some important aspects of the type of boundary migration connected with gradual grain growth. A distinctive feature of this type of migration is that the boundaries move *in the direction of their centers of curvature.* A typical example is shown in Fig. 1, which represents two successive stages in the migration of grain boundaries in an annealed high purity aluminum specimen. By means of a technique developed at this Laboratory,[4] the two successive stages are shown superimposed on each other. The driving energy of grain growth is the excess free energy associated with the grain boundary surfaces. Grain growth tends to decrease this free energy, thus allowing the structure to approach equilibrium, by decreasing the grain boundary curvatures and the total grain boundary surface per unit volume.

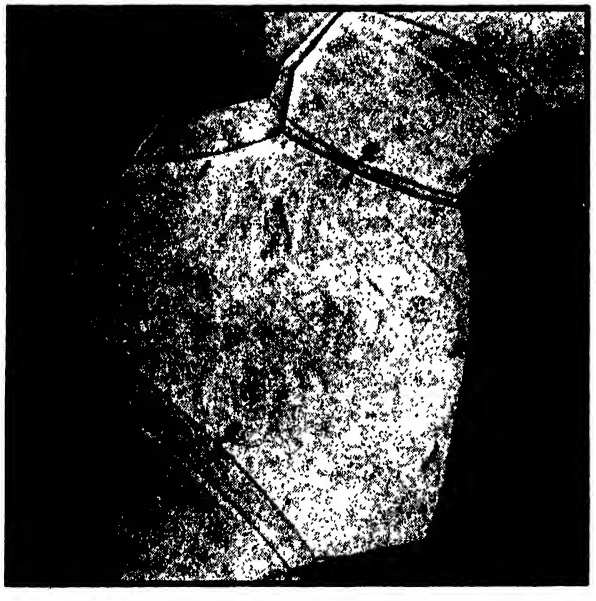


FIG. 1. In grain growth, boundary surfaces migrate toward their centers of curvature. Boundaries in high purity aluminum, after annealing 2 min. at 600°C (1), and after an additional anneal of 30 sec. at 600°C (2). Magnification 75×.

The recrystallization of cold worked metals also proceeds by means of boundary migration. However, in this case the moving boundaries separate strain hardened grains from annealed grains, and the direction of the movement is such that the volume of the annealed material increases with time at the expense of the cold worked material. The familiar concept of recrystallization involves the formation of nuclei of new strain-free grains in the midst of the strain hardened material. Most of the small new grains are at first entirely surrounded by one or two grains of the cold worked structure. During the growth of these new grains, their convex boundaries move in a direction *opposite to their centers of curvature.* Two stages of the growth of such a new grain (*C*) are shown in Fig. 2 by means of the same technique used in Fig. 1. The driving force of recrystallization is the excess free energy stored in the strain hardened material as a result of cold working. Even though the grain boundary movements here, as far as they can be observed under the microscope, take place in a direction opposite to that required by surface energy considerations, the total free energy is decreasing, and equilibrium is approached, as the strain hardened grains of high free energy content are replaced by low energy annealed grains.

Figure 2 also shows four successive stages of the migration of a grain boundary (1) already present in the original strain hardened structure. As revealed by the color effect in polarized light after electrolytic oxidation[4] and by x-ray diffraction, the orientation of the material swept by the moving boundary is very closely the same as that of the adjacent strain hardened grain (*A*). At first glance it might appear as if this boundary migration was related to ordinary grain growth, since one of the grains already present in the structure (*A*), is growing at the expense of another one

* Professor of Metallurgy, and Metallographer, respectively.

[1] D. Harker and E. R. Parker, "Grain shape and grain growth," Trans. A.S.M.E. 34, 156 (1945).
[2] C. S. Smith, "Grains, phases, and interfaces," Trans. A.I.M.E. 175, 15 (1948).
[3] J. E. Burke, "Some factors affecting the rate of grain growth in metals," Trans. A.I.M.E. 180, 73 (1949).
[4] Philip R. Sperry, Trans. A.I.M.E. 188, 103 (1949).

(*B*). On closer inspection, however, it becomes obvious that this process is different from grain growth, and that it is in important respects similar to recrystallization. This type of boundary migration takes place only in strain hardened material and it leads to an increase, rather than to a decrease, of grain boundary surface area, as seen in Fig. 2.

The question arises whether or not the observed boundary movements are motivated by the excess free energy of strain hardening, as in recrystallization. Figure 3 shows an example of the type of strain induced grain boundary migration discussed. The material swept by the moving boundary consists of a number of sections with very slightly different orientation. All these orientations are almost identical with the average orientation of the adjacent strain hardened grain (*A*). With the type of specimen preparation used, the surface of strain hardened grains becomes roughened, while the surface of strain free grains remains quite smooth. Figure 4 shows, by means of oblique illumination, that grains (*A*), (*B*) and (*C*), remaining from the original cold worked structure, have rough surfaces. On the other hand, the area swept by the moving boundary while migrating from position 1 to position 2, is quite smooth. As confirmed by means of x-ray diffraction, the material in this area is essentially in the annealed, strain-free, condition. It thus appears that in the strain induced boundary migration considered, as in ordinary recrystallization, the moving boundaries separate annealed material from strain hardened material and the direction of the boundary movement is such as to increase the volume of the annealed material. It is clear that the driving energy of the two processes is the same.

There is one significant difference, however, between the type of strain induced boundary migration shown

FIG. 2. In recrystallization (grains *C*, *D* and *E*) boundary surfaces migrate away from their centers of curvature. Strain induced growth of grain *A* at the expense of grain *B*. High purity aluminum annealed 90 min. at 600°C, rolled 7.5 percent (heavily outlined straight boundaries (1)), annealed successively 5 sec. (2), 5 sec. (3), 15 sec. (4), and 7 sec. (5) at 500°C. Magnification 75×.

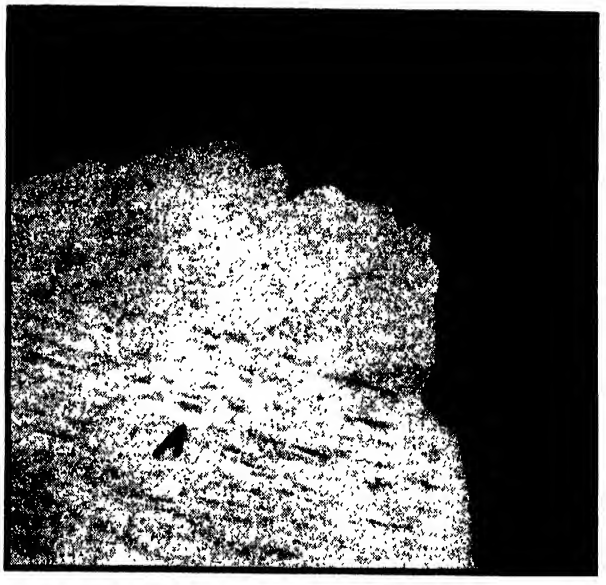


FIG. 3. Strain induced boundary migration in high purity aluminum. Orientation in area swept by migrating boundary is almost identical with that of strain hardened parent grain (A). Penultimate anneal: 5 min. at 650°C. 14 percent reduction by rolling. Last anneal 5 min. at 400°C. Magnification 75×.

in Fig. 3 and ordinary recrystallization. In the former, the annealed material swept by the migrating boundary has the same orientation as its strain hardened parent (*A* in Fig. 4), while in recrystallization a nucleus of different orientation is formed. As seen in Fig. 4, strain induced boundary migration without strain induced nucleation results in annealed areas bounded by an ordinary grain boundary on one side, and bordering on a strain hardened grain of the same orientation, without any intervening grain boundary, on the other side. That this is not so in ordinary recrystallization is shown in Fig. 5, where recrystallized grains *A* and *B* are located between strain hardened grains *C* and *D* and are separated from both of them by ordinary grain boundaries.

In summary, it may be stated that the processes of strain induced boundary migration found in aluminum is very much akin to recrystallization but that it differs from the latter by the fact that it is apparently nucleated by a strain hardened grain already present in the structure. Thus, *strain induced boundary migration may be considered as recrystallization without the formation of new nuclei.* Correspondingly, it was observed in many instances that, at the boundaries where it does occur, strain induced boundary migration starts promptly at the beginning of the annealing treatment. On the other hand, it is well known that a definite period of incubation is required for the formation of recrystallization nuclei and for their growth to microscopic sizes. This point is illustrated by Fig. 2, where each successive period of annealing has brought about a clearly observable migration of the boundary initially

FIG. 4. Strain induced boundary migration in high purity aluminum, as in Fig. 3. Strain hardened grains (*A*), (*B*), and (*C*) are roughened by etching, strain-free area, swept by migrating boundary, is smooth, and has no grain boundary toward grain (*A*). Oblique illumination.

FIG. 5. Recrystallized grains (*A*) and (*B*) are smooth and have grain boundaries toward both grains (*C*) and (*D*). The latter are strain hardened and consequently roughened by etching. Magnification 75×.

situated at (1). However, a recrystallized grain (*C*) came into existence only during the third period of annealing, and its two boundary positions, seen in the picture, correspond to the end of the last two annealing periods. Similarly, another recrystallized grain (*E*) was born during the fourth period of annealing and, therefore, the figure shows only one boundary position (5) for this grain. It was found that the ordinary recrystallization process, involving nucleation, plays a relatively minor role during the annealing of high purity aluminum deformed by rolling up to about 40 percent reduction of area. In such instances the type of strain induced boundary migration not involving the formation of new nuclei appears to predominate in the transformation to the strain-free condition.

This work was supported by the ONR, U. S. Navy, Contract No: N6 ori-165, T.O. 1.

The External Field Produced by a Slot in an Infinite Circular Cylinder*

SAMUEL SILVER AND WILLIAM K. SAUNDERS
Division of Electrical Engineering and the Antenna Laboratory, University of California, Berkeley, California
(Received August 10, 1949)

Expressions are derived for the external field produced by a slot of arbitrary shape in the wall of a circular wave guide (of infinite extent and infinite conductivity), the tangential components of the electric field in the slot being assumed to have been prescribed. This is accomplished by matching a Fourier representation of the external field, built up by superposition of basic sets of cylindrical waves, to a Fourier expansion of the prescribed field over the surface of the cylindrical wave guide. The far-zone field is obtained by applying the method of steepest descent to the Fourier integrals that constitute the coefficients in the expansion for the external field. The results satisfy the radiation conditions for far-zone fields.

I. INTRODUCTION

THIS paper, the first of a series of studies of the radiative properties of slots in conducting surfaces, is concerned with the field set up in space by a slot of arbitrary shape in the wall of a circular cylindrical wave guide. A complete analysis of the properties of a slot in the wall of a wave guide must make proper connection between the interior and exterior fields, thus determining the field distribution in the slot itself. Such a unified treatment is difficult to carry out and as yet has not been achieved for even the simplest slot configurations. Instead, the procedure has been to treat the interior and exterior problems separately in terms of assumed field distributions in the slot,[1] a procedure that we follow here. Under certain conditions, as for example in the case of an elongated narrow slot, reasonable assumptions can be made concerning the field distribution in the slot and results can be obtained that are of considerable usefulness in the design of radiating systems.

The subject of the slot in a circular wave guide has been treated previously under various restrictive conditions on its geometry and field distribution. G. Sinclair[2] and C. H. Papas and R. W. P. King[3] have discussed the field of a narrow slot of infinite length running parallel to the axis of the cylinder, excited uniformly along its length in the transverse direction. More recently C. H. Papas,[4] by means of a Green's function technique, has solved the problem of a finite slot in which the electric field has but a single tangential component that is parallel to the axis of the cylinder. This essentially restricts the configuration of the slot to be developable into a rectangle oriented with its boundaries respectively parallel to and perpendicular to the axis of the cylinder. As we shall show in Section V, the several previous results can be obtained directly from our general expressions.

II. GENERAL CONSIDERATIONS

The analysis presented here is free from the restrictive conditions on the slot geometry and fields that characterize the earlier work but is based also on the assumptions that the wave guide is infinite in extent and has infinite conductivity. The time dependence of the fields is taken to be harmonic and the complex exponential $e^{+i\omega t}$ is chosen as its representation.

The axis of the cylinder, which has a radius a, serves as the z axis of our coordinate system. Cylindrical coordinates ρ, ϕ, z are most suited for developing the general solution. Later, for purposes of describing the far-zone field, spherical coordinates R, θ, ϕ will be more appropriate; the cylinder axis will then form the polar axis of the spherical coordinate system. On the surface of the cylinder we have a slot of arbitrary shape, bounded in the axial direction by the planes $z=z_1$ and $z=z_2$ and defined circumferentially by the curves $\phi_1(z)$ and $\phi_2(z)$; the various quantities are shown in Fig. 1.

The tangential electric field in the slot will in general have both ϕ and z components which we consider to be prescribed functions $f_1(\phi, z)$ and $f_2(\phi, z)$, respectively. In view of the infinite conductivity of the wall

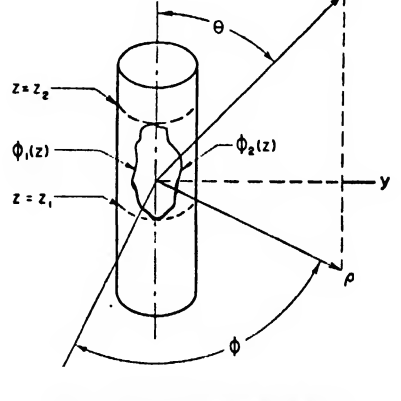


FIG. 1. Geometrical parameters.

* This work was supported in part by the Bureau of Ships, Contract NObsr-39401.

[1] For a general development of the interior problem see A. F. Stevenson, J. App. Phys. 19, 24 (1948).

[2] G. Sinclair, Proc. I.R.E. 36, 1487 (1948).

[3] C. H. Papas and R. W. P. King, Cruft Laboratory Technical Report No. 32, ONR Contract N5-ori-76 (March, 1948); Quart. App. Math. 7, No. 2 (July, 1949).

[4] C. H. Papas, Cruft Laboratory Technical Report No. 58, ONR Contract N5-ori-76 (September, 1948); J. App. Phys. 20, 437 (1949); Cruft Laboratory Technical Report No. 61, ONR Contract N5-ori-76 (October, 1948) to be published in J. Math. and Phys. 28, No. 4 (January, 1950).

this represents a complete knowledge of the tangential electric field over the entire surface of the cylinder, namely,

$$E_\phi = f_1(\phi, z) \quad \phi_1 \le \phi \le \phi_2;\ z_1 \le z \le z_2;$$
$$= 0 \qquad \text{outside the slot;} \tag{1a}$$

$$E_z = f_2(\phi, z) \quad \phi_1 \le \phi \le \phi_2;\ z_1 \le z \le z_2;$$
$$= 0 \qquad \text{outside the slot.} \tag{1b}$$

A solution of Maxwell's equations must be constructed that assumes the prescribed value of the tangential electric field over the cylinder and that satisfies certain radiation conditions at infinity. For a time-periodic field the latter take the form:[5]

$$\lim_{R\to\infty} R\mathbf{E} \text{ is finite,} \tag{2a}$$

$$\lim_{R\to\infty} R[\mathbf{R}_1 \times \mathbf{H} + (\epsilon/\mu)^{\frac{1}{2}}\mathbf{E}] = 0, \tag{2b}$$

$$\lim_{R\to\infty} R\mathbf{H} \text{ is finite,} \tag{2c}$$

$$\lim_{R\to\infty} R[(\mathbf{R}_1 \times \mathbf{E}) - (\mu/\epsilon)^{\frac{1}{2}}\mathbf{H}] = 0, \tag{2d}$$

where $\mathbf{R}_1$ is a radial unit vector in the spherical coordinate system (see Fig. 1). Equations (2b) and (2d) are the radiation conditions proper and express the fact that we have only outgoing waves at infinity. A time-periodic solution that meets the above requirements is unique.

The procedure is to construct Fourier representations for the tangential electric field over the cylinder and for the field in space. The latter is synthesized by superposition of basic sets of cylindrical waves that satisfy the requisite conditions at infinity. The coefficients of the expansion for the surface field are known; the coefficients of the representation for the space field are determined by the requirement that its tangential components at the surface of the cylinder shall reduce to the Fourier representation of the field prescribed over the latter. The resulting expressions are in the form of a Fourier series whose coefficients are Fourier integrals.

III. DEVELOPMENT OF THE FIELDS

The Tangential Field over the Cylinder

We shall first develop the Fourier expansion for the tangential components of the electric field over the surface of the cylinder. Let f_α denote either $f_1(\phi, z)$ or $f_2(\phi, z)$ and let E_α be the corresponding field component. In the ϕ-direction E_α is a periodic function and is therefore to be represented by a Fourier series

$$E_\alpha(\phi, z) = \sum_{n=-\infty}^{\infty} A_n{}^\alpha(z) e^{-jn\phi}, \tag{3}$$

the coefficients $A_n{}^\alpha$ being functions of z. It is readily evident that

$$A_n{}^\alpha = \frac{1}{2\pi}\int_{\phi_1(z)}^{\phi_2(z)} f_\alpha(\beta, z) e^{jn\beta} d\beta, \quad z_1 \le z \le z_2; \tag{4a}$$

$$= 0. \qquad z < z_1, \quad z > z_2. \tag{4b}$$

$A_n{}^\alpha(z)$ is thus a piecewise continuous function and its Fourier representation is the Fourier integral

$$A_n{}^\alpha = \frac{1}{2\pi}\int_{-\infty}^{\infty} dh \int_{z_1}^{z_2} A_n{}^\alpha(\xi) e^{-jh(z-\xi)} d\xi; \tag{5}$$

or,

$$A_n{}^\alpha(z) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} dh \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_\alpha(\beta, \xi) e^{jn\beta} e^{-jh(z-\xi)} d\beta. \tag{6}$$

The tangential electric field over the cylinder is thus expressed in the form of a Fourier expansion:

$$E_\alpha(\phi, z) = \frac{1}{4\pi^2}\sum_{n=-\infty}^{\infty} e^{-jn\phi} \int_{-\infty}^{\infty} dh \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_\alpha(\beta, \xi) e^{jn\beta} e^{-jh(z-\xi)} d\beta. \tag{7}$$

The External Field Expansion

We construct the external field by superposition of basic sets of cylindrical waves. The reader is referred to Stratton[6] for an exposition of the general theory of the latter. The basic sets fall into two groups: the transverse magnetic (TM) type having no H_z component and the transverse electric (TE) type having no E_z component. Appropriate combinations of these can be made to form a basic set of field components from which a general field that is periodic in ϕ can be synthesized. Each member of the latter set is a Fourier series as follows:

$$\mathcal{E}_\rho = \left\{ \sum_{n=-\infty}^{\infty} \left[-jha_n \frac{\partial H_n{}^{(2)}(\Lambda\rho)}{\partial\rho} - \frac{n\omega\mu}{\rho} b_n H_n{}^{(2)}(\Lambda\rho) \right] e^{-jn\phi} \right\} e^{-jhz}, \tag{8a}$$

$$\mathcal{E}_\phi = \left\{ \sum_{n=-\infty}^{\infty} \left[-\frac{nh}{\rho} a_n H_n{}^{(2)}(\Lambda\rho) + j\omega\mu b_n \frac{\partial H_n{}^{(2)}(\Lambda\rho)}{\partial\rho} \right] e^{-jn\phi} \right\} e^{-jhz}, \tag{8b}$$

[5] S. Silver, *Microwave Antenna Theory and Design* (McGraw-Hill Book Company, Inc., New York, 1949), Radiation Laboratory Series, Vol. 12, Chapter 3.

[6] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), Chapter 6. Note that our expressions differ from those of Stratton in signs because of our choice of $e^{+j\omega t}$ rather than $e^{-j\omega t}$ to represent the time dependence. This also leads to the use of the Hankel function of the second kind to represent an outward traveling wave.

$$\mathcal{E}_z=\left\{\sum_{n=-\infty}^{\infty}[(k^2-h^2)a_nH_n^{(2)}(\Lambda\rho)]e^{-jn\phi}\right\}e^{-jhz}, \quad (8c)$$

$$\mathcal{H}_\rho=\left\{\sum_{n=-\infty}^{\infty}\left[\frac{nk^2}{\mu\omega\rho}a_nH_n^{(2)}(\Lambda\rho)-jhb_n\frac{\partial H_n^{(2)}(\Lambda\rho)}{\partial\rho}\right]e^{-jn\phi}\right\}e^{-jhz}, \quad (8d)$$

$$\mathcal{H}_\phi=\left\{\sum_{n=-\infty}^{\infty}\left[-j\frac{k^2}{\mu\omega}a_n\frac{\partial H_n^{(2)}(\Lambda\rho)}{\partial\rho}-\frac{nh}{\rho}b_nH_n^{(2)}(\Lambda\rho)\right]e^{-jn\phi}\right\}e^{-jhz}, \quad (8e)$$

$$\mathcal{H}_z=\left\{\sum_{n=-\infty}^{\infty}[(k^2-h^2)b_nH_n^{(2)}(\Lambda\rho)]e^{-jn\phi}\right\}e^{-jhz}, \quad (8f)$$

with

$$\Lambda=[k^2-h^2]^{\frac{1}{2}}. \quad (9)$$

In the case of harmonic time dependence k is a constant that is given in terms of the constitutive parameters—magnetic permeability μ, electric permittivity ϵ, conductivity σ—of the medium under consideration and the angular frequency by

$$k=[\omega^2\mu\epsilon-j\mu\sigma\omega]^{\frac{1}{2}}; \quad -\pi/2<ph\,k\leq 0. \quad (10)$$

The parameter h with respect to which the set is ordered is in general complex; its values are to be taken to suit the required boundary conditions. The Hankel function of the second kind enters into the wave functions to represent outgoing waves. However, this property of the Hankel functions depends on the definition of the phase of Λ for the various values of h. We shall return to this matter shortly.

The terms in Eqs. (8) having the common coefficients a_n belong to the set of TM waves; those having common coefficients b_n are of the TE type. It will be observed that there are only two sets of independent coefficients and hence, a knowledge of any two field components is sufficient to establish the complete field. The coefficients a_n and b_n are independent of ρ and z but in general are functions of h.

We can now build up an arbitrary z dependence by superposition of components of various values of h. Regarding h as a complex quantity, different representations can be obtained according to the path in the h plane along which h is allowed to vary. In the present case where we are to fit the field at $\rho=a$ to a Fourier integral along the surface of the cylinder the path along the real-h axis is indicated. Thus for any field component $E_\alpha(\rho, \phi, z)$ we shall write

$$E_\alpha(\rho, \phi, z)=\int_{-\infty}^{\infty}\mathcal{E}_\alpha(\rho, \phi, h)dh. \quad (11)$$

Similar expressions hold for the magnetic field components.

We have noted before that the phase of $\Lambda=(k^2-h^2)^{\frac{1}{2}}$ must be defined along the path of integration in such a way as to produce proper behavior of the field at infinity. The locations of the branch points $h=\pm k$ are shown in Figs. 2(a) and 2(b), the latter applying to the case when the external medium has neither an ohmic nor a dielectric dissipation factor. To define the phase of Λ let us consider the behavior of $H_n^{(2)}(\Lambda\rho)$ for very large values of ρ, making use of the asymptotic form of the Hankel function,

$$H_n^{(2)}(\Lambda\rho)\approx[2/\pi\Lambda\rho]^{\frac{1}{2}}e^{-j\Lambda\rho}e^{j(2n+1)\pi/4}. \quad (12)$$

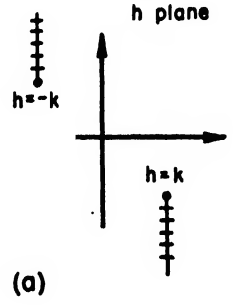
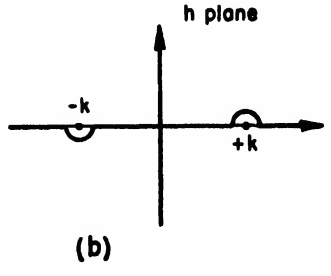


FIG. 2. Path of integration in the h Plane: (a) k complex; (b) k real.



For the general case, when k is complex, it is seen that it is necessary to require

$$Re\Lambda\geq 0, \quad Im\Lambda\leq 0 \quad (13)$$

in order to have outward radial propagation and to keep the field components bounded at infinity. If k is real, the conditions (13) reduce to

$$ph\Lambda=0 \qquad |h|<k, \quad (14a)$$

$$ph\Lambda=-\pi/2 \qquad |h|>k, \quad (14b)$$

and it is seen that the contour must be indented around the branch points in the manner shown in Fig. 2(b).

Evaluation of the Coefficients

With the sense of Λ properly defined the components of the external field are given by Eq. (11). To evaluate

the coefficients a_n and b_n we match the components E_ϕ and E_z to the assigned values over the cylinder. From Eqs. (8) and (11), on interchanging summation and integration, we obtain

$$E_\phi = \sum_{n=-\infty}^{\infty} e^{-jn\phi} \int_{-\infty}^{\infty} \left[-\frac{nh}{\rho} a_n H_n^{(2)}(\Lambda\rho) + j\omega\mu b_n \frac{\partial H_n^{(2)}(\Lambda\rho)}{\partial\rho} \right] e^{-jhz} dh, \quad (15a)$$

$$E_z = \sum_{n=-\infty}^{\infty} e^{-jn\phi} \int_{-\infty}^{\infty} (k^2 - h^2) a_n H_n^{(2)}(\Lambda\rho) e^{-jhz} dh. \quad (15b)$$

At $\rho = a$ these must reduce to the expansions of Eq. (7); hence,

$$\left[-\frac{nh}{a} H_n^{(2)}(\Lambda a) \right] a_n + \left[j\omega\mu \frac{\partial H_n^{(2)}(\Lambda a)}{\partial a} \right] b_n = \frac{1}{4\pi^2} \int_{z_1}^{z_2} d\xi \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_1(\beta, \xi) e^{jn\beta} e^{jh\xi} d\beta, \quad (16a)$$

$$[(k^2 - h^2) H_n^{(2)}(\Lambda a)] a_n = \frac{1}{4\pi^2} \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi) e^{jn\beta} e^{jh\xi} d\beta; \quad (16b)$$

or,

$$a_n = \frac{1}{4\pi^2 (k^2 - h^2) H_n^{(2)}(\Lambda a)} \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi) e^{jn\beta} e^{jh\xi} d\beta, \quad (17a)$$

$$b_n = \frac{-jnh}{4\pi^2 \omega\mu a (k^2 - h^2) (\partial H_n^{(2)}(\Lambda a)/\partial a)} \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi) e^{jn\beta} e^{jh\xi} d\beta - \frac{j}{4\pi^2 \omega\mu (\partial H_n^{(2)}(\Lambda a)/\partial a)} \int_{z_1}^{z_2} d\xi \times \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_1(\beta, \xi) e^{jn\beta} e^{jh\xi} d\beta. \quad (17b)$$

These taken together with Eqs. (8) and (11) give the formal solution to the problem. The surface current distribution **K** over the cylinder is given directly by the tangential magnetic field at the surface:

$$K_\phi = -H_z(a, \phi, z); \quad K_z = H_\phi(a, \phi, z). \quad (18)$$

Whereas the process of obtaining the formal solution is a relatively simple one, the numerical evaluation of the fields is a task of considerable magnitude, particularly for points in the neighborhood of the slot. However, for the far-zone field the expressions simplify to a large extent and numerical work becomes more tractable. We shall discuss this region of the field in the following section.

IV. THE FAR-ZONE FIELD

The reduction of the general expressions for this region is effected by making use of the asymptotic forms of the Hankel functions and subsequently carrying out certain integrations by the saddle-point method.[8] We shall restrict the treatment to the case when k is real.

It is convenient to describe the far-zone field in terms of spherical coordinates R, θ, ϕ (see Fig. 1), and, accordingly we make the substitutions

$$\rho = R\sin\theta, \quad z = R\cos\theta \quad (19)$$

in the expressions for the components of the field. We are interested in the dominant characteristics of the field at large distances R; that is, we wish to obtain the leading terms in a development of the field components in inverse powers of R. Except for small values of θ we may, for large values of R, replace $H_n^{(2)}(\Lambda\rho) = H_n^{(2)}(\Lambda R\sin\theta)$ by the asymptotic form given in Eq. (12).† When this is done and the order of integration is suitably rearranged, it is found that the expressions for the field components take the general form

$$\frac{1}{4\pi^2} \sum_{n=-\infty}^{\infty}{}' e^{-jn\phi} \left\{ \int_{z_1}^{z_2} d\xi \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_1(\beta, \xi) e^{jn\beta} d\beta \times \int_{-\infty}^{\infty} u(h, R, \theta; n) e^{jh\xi} e^{-jR[\Lambda \sin\theta + h\cos\theta]} dh + \int_{z_1}^{z_2} d\xi \int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi) e^{jn\beta} d\beta \times \int_{-\infty}^{\infty} v(h, R, \theta; n) e^{jh\xi} e^{-jR[\Lambda \sin\theta + h\cos\theta]} dh \right\}. \quad (20)$$

The functions u and v can readily be evaluated for each particular component. It will be observed that they vary slowly over the range of h as compared with the exponential functions. The integrals over h in (20) are, therefore, of the general type to which the saddle-point method is applicable. We are here primarily concerned with the region of the field where R is very

[8] See, for example, G. N. Watson, *Bessel Functions* (Cambridge University Press, London, 1944), pp. 235 ff.

† We shall not enter here into a consideration of the different asymptotic representations that must be used according as to whether the order of the Hankel function is small, nearly equal to, or large compared with the magnitude of $\Lambda\rho$. From a practical standpoint these details are not too important inasmuch as the terms of the series are found to diminish rapidly with increasing values of n.

much larger than both the wave-length and the axial extent of the slot ($R \gg \xi$). The variation of the phase factor involving R thus dominates that involving ξ and we have the simplest form of saddle-point problem. It will be well to note in passing that in the case of a slot whose axial extent is large compared with the wave-length there is an intermediate zone, at distances about equal to or possibly less than the total axial dimension, for which (20) also gives the dominant terms. In the latter case, however, we have two rapidly varying phase factors to consider in the integration over h and consequently a saddle-point problem involving two large parameters.‡ Inasmuch as the intermediate zone

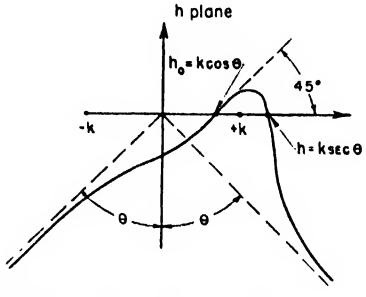


FIG. 3. Modified path through the saddle-point.

is of limited interest it will not be discussed further in this paper. Practical interest usually centers about the far-zone field pattern which is one of the primary specifications that are laid down for the design of a given antenna system.

With the restriction $R \gg \xi$ we may formulate our problem as that of evaluating an integral of the form

$$\int_{-\infty}^{\infty} F(h)e^{-jR[\Lambda \sin\theta + h\cos\theta]}dh = \int_{-\infty}^{\infty} F(h)e^{R\Phi(h)}dh \quad (21)$$

for large values of R, with

$$F(h) = \left.\begin{matrix} u(h, R, \theta; n) \\ v(h, R, \theta; n) \end{matrix}\right\} e^{jh\xi}. \quad (21a)$$

The saddle-point technique consists of deforming the path of integration into one that passes through the point h_0 for which $\Phi'(h_0) = 0$ and is such that over its entirety

$$Im\Phi(h) = Im\Phi(h_0).$$

If the integral converges along this path, the major contribution to the integral arises from a small segment about the saddle-point $h = h_0$. In our case the new path is a function of θ; its general form is shown in Fig. 3. The deformation from the real axis to this path is possible by virtue of the analytic properties of $F(h)$ and by the extension of conditions such as were imposed on Λ in Eq. (13) to the entire exponent, namely $Im\Phi(h) < 0$. This makes the integral over the arc connecting the real axis and the new path at infinity vanish, and also ensures the convergence of the integral along the modified path. Considering the small segment of the latter in the neighborhood of h_0 having the modulus $u < \epsilon$, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} F(h)e^{R\Phi(h)}dh \\
&\quad \approx F(h_0)e^{[R\Phi(h_0)+j\pi/4]} \int_{-\epsilon}^{\epsilon} e^{j\Phi''(h_0)u^2/2}du \\
&\quad = F(k\cos\theta)e^{-j(kR-\pi/4)} \int_{-\epsilon}^{\epsilon} e^{-Ru^2/2k\sin^2\theta}du \\
&\quad \approx F(k\cos\theta)e^{-j(kR-\pi/4)}\left[\frac{2k\sin^2\theta}{R}\right]^{\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\tau^2}d\tau;
\end{aligned}$$

or,

$$\int_{-\infty}^{\infty} F(h)e^{R\Phi(h)}dh \approx \left[\frac{2\pi k}{R}\right]^{\frac{1}{2}} \sin\theta F(k\cos\theta)e^{-jkR}e^{j\pi/4}. \quad (22)$$

The leading terms are found to be of order R^{-1}. These form the radiation field. The components of E_ρ and E_z in the R-direction that are of this order of magnitude cancel one another; the same occurs in the magnetic field. The radiation field is thus transverse to the R-direction and, in general, has non-zero components:

$$E_\theta = -\frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\sum_{n=-\infty}^{\infty}\frac{j^{n+1}e^{-jn\phi}}{\sin\theta H_n^{(2)}(ka\sin\theta)} \times \int_{z_1}^{z_2} d\xi e^{jk\xi\cos\theta}\int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi)e^{jn\beta}d\beta; \quad (23a)$$

$$\begin{aligned}
E_\phi = \frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\sum_{n=-\infty}^{\infty}\frac{j^n e^{-jn\phi}}{H_n^{(2)\prime}(ka\sin\theta)} \\
\times\left[\int_{z_1}^{z_2} d\xi e^{jk\xi\cos\theta}\int_{\phi_1(\xi)}^{\phi_2(\xi)} f_1(\beta, \xi)e^{jn\beta}d\beta\right. \\
\left. + \frac{n\cot\theta}{ka\sin\theta}\int_{z_1}^{z_2} d\xi e^{jk\xi\cos\theta} \times\int_{\phi_1(\xi)}^{\phi_2(\xi)} f_2(\beta, \xi)e^{jn\beta}d\beta\right]; \quad (23b)
\end{aligned}$$

$$H_\theta = -(\epsilon/\mu)^{\frac{1}{2}}E_\phi; \quad H_\phi = (\epsilon/\mu)^{\frac{1}{2}}E_\theta. \quad (23c)$$

These show explicitly that the solutions satisfy the required conditions at infinity that are formulated in Eqs. (2).

‡ The problem is very similar to that of obtaining various asymptotic forms of the cylinder functions according to the relative magnitudes of the order and the argument of the functions; see reference (8), Chapter 8.

V. ILLUSTRATIVE EXAMPLES

To illustrate the applicability of the results we shall touch briefly upon two cases reported previously in the literature to show how the earlier results are readily obtained from our general expressions:

1. The Uniformly Excited Circumferential Slot[4]

The slot runs completely around the cylinder and has a height $2w$ that is very much smaller than the circumference. For convenience we shall take $z_1=-w$, $z_2=w$. Note also that $\phi_1(z)=0$, $\phi_2(z)=2\pi$. The tangential electric field in the slot has only a z component and that independent of ϕ; thus

$$f_1(\phi, z)=0; \quad f_2(\phi, z)=E_0. \tag{24}$$

The integrals over β in Eqs. (17) vanish except for $n=0$. Consequently, the only non-zero coefficient is

$$a_0=\frac{2E_0w(\sin hw/hw)}{2\pi(k^2-h^2)H_0^{(2)}(\Lambda a)}. \tag{25}$$

Defining the voltage across the slot to be $V=2E_0w$, we get for the general expression for the z component of the field in space

$$E_z(\rho, \phi, z)=\frac{V}{2\pi}\int_{-\infty}^{\infty}\frac{H_0^{(2)}[(k^2-h^2)^{\frac{1}{2}}\rho]}{H_0^{(2)}[(k^2-h^2)^{\frac{1}{2}}a]}\frac{\sin hw}{hw}e^{-jhz}dh. \tag{26}$$

If the slot is a δ-slot, that is, $w\to 0$ while V is constant, $\sin hw/hw$ reduces to unity. The far-zone field is also readily obtained by evaluating the $n=0$ term of Eq. (23a) for the given $f_2(\phi, z)$.

2. The Infinite Axial Slot[2,3]

Another case that has been dealt with in the literature is that of a narrow rectangular slot, parallel to the axis of the cylinder, over which the field has only an E_ϕ component. The width of the slot will be taken to be $2a\phi_0$. The assumed field distribution in the slot is

$$f_1(\phi, z)=E_0, \quad f_2(\phi, z)=0. \tag{27}$$

All the coefficients a_n are now equal to zero, while for b_n we get

$$b_n=\lim_{l\to\infty}\frac{E_0}{j4\pi^2\omega\mu(\partial H_n^{(2)}(\Lambda a)/\partial a)}\times\int_{-l}e^{jh\xi}d\xi\int_{-\phi_0}e^{jn\beta}d\beta. \tag{28}$$

Defining a voltage $V=2E_0a\phi_0$ and making use of the Fourier integral definition of the δ-function,

$$\delta(h-0)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{jh\xi}d\xi \tag{29}$$

we obtain finally for the E_ϕ component of the field

$$E_\phi=\frac{V}{2\pi a}\sum_{n=-\infty}^{\infty}\frac{\sin n\phi_0}{n\phi_0}\frac{H_n'^{(2)}(k\rho)}{H_n'^{(2)}(ka)}e^{-jn\phi}. \tag{30}$$

If the slot is regarded as a δ-slot in the ϕ-direction, $\sin n\phi_0/n\phi_0\to 1$ for all n.

Further application of the theory will be presented in a forthcoming paper in which we discuss the transverse rectangular slot in detail and give experimental data in conjunction with the theoretical results.

The Reversal Theorem of Linearized Supersonic Airfoil Theory*

M. M. Munk
Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland
(Received May 27, 1949)

It is demonstrated on simple dynamic principles that a supersonic airfoil under certain conditions exhibits the same drag and the same lift slope regardless of whether it flies forward or backward.

THE thin airfoil theory in the supersonic regime is ordinarily referred to as the linearized theory. The boundary condition is simplified exactly as in the case of the incompressible fluid. In addition, the partial differential equation is also simplified, and it reduces to the wave equation:

$$\frac{\partial^2 F}{\partial x^2}+\frac{\partial^2 F}{\partial y^2}-\frac{\partial^2 F}{\partial z^2}=0. \tag{1}$$

Herein, F denotes either the velocity potential, any one Cartesian velocity component, or the pressure. The variables x, y, and z denote Cartesian coordinates in spanwise, liftwise, and fore-aft-wise direction of an airfoil, respectively. The corresponding components of the increment velocity are u, v, and w. The increment velocity is small when compared with the velocity of advance W_0.

In the computations of this theory, the density ρ_0 occurs as if it were constant. Absence of divergence of velocity is not assumed, not even approximately, in conjunction with such constant density. The increment pressure Δp is

$$\Delta p=-\rho_0 w W_0. \tag{2}$$

The rate at which energy is left behind the airfoil advancing with the velocity W_0 in air otherwise at rest is equal to DW_0, the product of the drag and of the velocity of advance. We assume specifically that the drag as defined from this relation is consistent with what follows, especially with the absence of a finite chordwise, leading-edge force. Since Eq. (1) is linear, the superposition of any number of individual solutions will again be a solution.

The vortex theorems hold. All vortices within the air are parallel to the Z direction. No finite force acting on an infinitely thin leading edge in the chordwise direction is supposed to exist. The opposite would be inconsistent with finite bounds for the pressure or suction. The resultant airfoil force is, therefore, merely the sum total of the ordinary pressure effects. With a constant local angle of attack of a plane, infinitely thin airfoil, the resultant airforce is at right angles to the airfoil. The constant angle of attack is equal to the drag-lift ratio. Both the leading edge and the trailing edge are points or lines of separation or rejoining of the flow. It is assumed that this double-edge condition is not inconsistent with the existence of a solution of Eq. (1). The solution is not made unique by selecting the trailing edge as a rejoining line. Rather, the solution is made unique by stipulating that all free vortices be trailing behind the airfoil and that all pressure waves proceed in the rear of the airfoil. In other words, all perturbations are restricted to the region downstream of the leading edge of the airfoil. With a large sweepback angle such as occurs with arrow wings, there may not exist any solution of (1) if both the leading separation line and the trailing separations line are arbitrarily specified. This is the case of the "subsonic" leading edge, where the downstream Mach cones, having their apices at the trailing edge, intersect with or enclose the leading edge. In that case it is necessary that the leading separation line be compatible with the other geometric conditions including the position of the trailing separation line, that is the trailing edge and with Eq. (1).

There exists an infinity of solutions that do not comply with this last stipulation. All of these solutions are acceptable on physical grounds and occasionally represent what happens or what may happen.

Reversing all velocities or all increment velocities leads again to a solution of (1).

Equation (1) is the basic equation in acoustics. Therefore, it has been stated that acoustical problems are analogous to the airfoil problems now considered. That is an extreme understatement: the linearized airfoil problem is itself an acoustical problem. The solutions exhibit all ordinary features of acoustical phenomena, wave effects, focusing effects, and that pronounced absence of harmony and of the continuity of analytic functions peculiar to acoustics. The solutions are disharmonic and non-unitary—a patchwork of heterogeneous pressure distributions of contrasting character. The computation has to be adapted to this state of things. One region after another region has to be computed in piecemeal fashion. No individual solution obtained may be applied to a slightly modified airfoil or Mach number by the application of the intuitive faculty along without careful scrutiny.

Notwithstanding their erratic character, the solutions are subordinated to simple, broad, and universal principles, transcending the irregularity in detail. Consider for instance a plane strip of constant width b extending streamwise into infinity. Its lift is finite, not even depending on the Mach number, such that the angle of

* Sponsored by the Office of Naval Research.

attack is equal to

$$\alpha = \frac{2L}{\rho_0 W_0^2 b^2 \pi}. \tag{3}$$

The spanwise lift distribution approaches (summed up streamwise) the elliptic lift distribution. The lift density asymptotically approaches zero in downstream direction, but it approaches it in a kind of periodic fashion and by no means monotonically.

If the strip is cambered with straight parallel span lines, the angle of attack at the leading edge is zero, the streamwise curvature also approaches zero and the angle of attack approaches asymptotically the above value (3). Then the drag will assume the minimum induced drag of the incompressible fluid theory. This is a very academic point, indeed, for the large airfoil area required would lead to excessive and dominating surface friction drag forces. Still, this illuminates the conditions and exemplifies the simple relations transcending the erratic details.

Many individual solutions of (1) relating to airfoils of various planviews and of wind sections have been determined. Those working in this field observed that all solutions are consistent with a remarkable symmetry principle. [See Bibliography.] Any airfoil moved under otherwise equal conditions forward and backward yields equal drag in these linearized computations. Turning the airfoil by 180° about a spanwise direction changes in general the lift distribution and the drag distributions, and sometimes very radically too. But the computed sum total of the drag remain always the same. With a plane, infinitely thin airfoil the computed lift also remains the same, so that for any airfoil the computed lift slope remains the same.

We proceed now to present a general proof of this reversal theorem based on very simple grounds. Picture the airfoil in place and expose it to the undisturbed flow (A). An increment flow (a) in rear of the airfoil is then produced. Leave the airfoil in place, but expose it now to an undisturbed airflow (B) having equal but opposite velocity as (A). This time an increment flow (b) will be created downstream of the airfoil in relation to (B), but upstream of the airfoil in relation to (A). The two increment flows (a) and (b) are ordinarily quite different, but not wholly uncorrelated. Combine or superpose now the flows (A), (a), and (b). This combination flow is now associated with free vortices and with pressure waves upstream of the airfoil as well as with free vortices and with pressure waves downstream of the airfoil. Right at the airfoil surface, all liftwise increment velocities at right angles to the airfoil cancel out. The airfoil consistent with the superposition flow is, therefore, plane, infinitely thin, and has zero incidence. The airfoil still has a lift distribution, but it cannot have any drag. The rate of energy transport upstream of the airfoil is equal to that of the flow (B)–(b). The rate of energy transport downstream of the airfoil is that of the flow (A)–(a). Both are alike, because the drag of the combination airfoil is zero. Hence the drag of the reversed airfoil must be equal to the drag of the airfoil in its initial direction.

Since, furthermore, the angle of attack of the plane infinitely thin airfoil does not change by the reversal, and since this angle is equal to the drag-lift ratio, it follows that for such special airfoils having a drag the total lift in both cases must also be the same. Hence, the lift slope must remain unchanged by the reversal for any airfoil.

Two airfoils may, of course, be associated with equal drag and with equal lift slope without being the reversal of each other. If the airfoils are the reversal of each other, no further proof is needed for equal drag and lift slope. The airfoils are not truly reversals of each other if the leading edge of the profile remains the leading edge unless the profiles are symmetric fore and aft. A full bodily reversal of the airfoil is necessary to make the preceding proof applicable. It is, on the other hand, apparent that the proof does not call for a single airfoil. Any assembly or combination of several airfoils in multiplane or in tandem relations or both, is permitted without invalidating the preceding proof. The airfoils must then not be reversed individually, but in the whole formation.

The proof does not lead to any information about the center of pressure of the reversed airfoil to its lift, unless the airfoil has zero camber. The symmetric relation holds only for the lift slope of an airfoil with arbitrary camber.

The argument of paragraph 10 implies the requirement that both the leading edge and the trailing edge be defined and serve as the separation line and as the rejoining line of the flow, respectively. Otherwise, very large streamwise increment velocity components would be generated, in contradiction with the assumptions made leading to the adoption of the linearized flow. The streamwise increment velocity component would not then be small throughout when compared with the velocity of advance. Hence the cancellation of the liftwise increment-velocity components of (a) and (b) in paragraph 9 would not be assured. Also the basic laws of mechanics would not be approximated by a solution obtained from the linearized differential equation. It is also implicit that the existence of a solution is not inconsistent with one double-edge condition. The requirement regarding the leading edge and the trailing edge is inconsistent with the existence of a solution of Laplace's equation in the theory of airfoils moving through an incompressible fluid. The requirement may be inconsistent with the existence of a solution of Eq. (1). It is by no means always inconsistent. It is not if the leading and the trailing edge are both "supersonic" throughout, that is, if the edges form an angle with the direction of advance larger than the

Mach angle. It is also not inconsistent if the local mean angle of attack of the airfoil is zero at all points. Supersonic edges throughout are not a necessary condition for the existence of a solution. There is some indication that nonexistence of a solution may require that portions of the leading edge be downstream and in the Mach direction relative to portions of the trailing edge. It seems that it has not yet been determined rigorously or finally under what conditions the two edges may be prescribed to serve as separate line and rejoining line, respectively, without excluding a solution of Eq. (1), and under what conditions they may not.

BIBLIOGRAPHY

L. Prandtl, "Theorie des Flugzeugtragflügels im zusammendrückbaren Medium," Luftfahrt-Forschung 13, 313–319 (1936).

P. A. Lagerstrom, "Formulas in Three Dimensional Wing Theory," Douglas Aircraft Company, Report No. SM 11901 (July 8, 1946).

Th. von Karman, "Supersonic Aerodynamics," J. Aero. Sci. 14, 373–402 (1947). See Section 5d.

The Impact of a Body on a Water Surface at an Arbitrary Angle

LEON TRILLING*

Hydrodynamics Laboratory, California Insitute of Technology, Pasadena, California

(Received August 18, 1949)

This paper presents an approximate method of determining the pressure distribution during impact on the surface of a body which strikes a horizontal water surface at an arbitrary angle.

The effect of the splash is neglected and the pressure on the free boundary is assumed proportional to the potential, as if the process were an impulse. The shape of the submerged portion of the striking body is approximated by a semi-ellipse (two dimensions), by a hemisphere, half an ellipsoid of revolution and half a general ellipsoid. Under those conditions, explicit results for the pressure distribution are found. As a special example, the impact pressures on a sphere striking at 45° angle are computed in detail.

AN estimate of the pressure forces on a body which strikes a water surface at an arbitrary angle is important in many problems, such as the design of seaplane hulls or naval ballistics.

The present work discusses the earliest stage of the water entry phenomenon which precedes the appearance of a cavity and displays the highest pressures. It is usually called the impact stage.

The problem of water impact has been discussed by numerous writers interested in the landing characteristics of seaplanes and the underwater trajectory of projectiles. Because of the difficulty of the problem, they were forced to various simplifying assumptions. Thus, von Karman[1] approximated the shape of the striking body by a growing flat plate in the plane of the water surface. H. Wagner[2,3] refined Karman's analysis by making some allowance for the water splash, and M. S. Plesset[4] by replacing the flat plate by an elliptic disk. To take into account the penetration of the body into the water, Schiffman and Spencer[5,6] approximated its shape by a spherical lens; P. Y. Chou[7] approximated it by a spherical bowl, whose potential was calculated by A. B. Basset.[8] Those latter investigations give accurate results in the case of normal impact, but they are too complicated to be generalized to impact at an arbitrary angle. Some simple two-dimensional problems were solved in all generality by L. I. Sedoff[9] who used the powerful methods of conformal transformation.

In the following pages, the problem is linearized and the classical methods of potential theory are applied to determine the pressure on two-dimensional bodies whose submerged portion may be approximated by a semi-ellipse, and on three-dimensional bodies which may be approximated by a hemisphere, half an ellipsoid of revolution, and half a general ellipsoid. Because of the linear nature of the simplified problem the effect of the vertical and horizontal components of impact velocity may be discussed separately and superimposed, so that one obtains results valid for any angle of impact.

1. STATEMENT OF THE PROBLEM

A three-dimensional half-space filled with incompressible fluid of density ρ at rest with a horizontal free

* The major portion of this study was carried out while the writer was with the Naval Ordnance Test Station; the writer also wishes to acknowledge the support of the ONR.

[1] Th. von Karman, "The impact on sea-plane floats during landing," NACA-TN 321 (1929).

[2] H. Wagner, "On the landing of seaplanes," ZFM (January, 1931)—NACA-TM 622.

[3] H. Wagner, "Uber stoss und gleitvorgange an der oberflache von flussigkeiten," ZAMM, 12, 193–215 (1932).

[4] M. S. Plesset, "An investigation of the impact forces on torpedoes entering water," Douglas Company Report SM3937 (December, 1942).

[5] Schiffman and Spencer, "The force of impact on a sphere striking a water surface," AMP-421R-AMG-NYU No. 105 (February, 1945).

[6] Schiffman and Spencer, "The force of impact on a sphere striking a water surface (2)," AMP-42.2R-AMG-NYU No. 133 (July, 1945).

[7] P. Y. Chou, "On impact of spheres upon water," U.S. NOTS (1946), (to be published).

[8] A. B. Basset, "On the potential of an electrified spherical bowl and on the velocity potential due to the motion of an infinite liquid about such a bowl," Proc. Math. Soc. London, 1st Series 16, 286–306 (1885).

[9] L. I. Sedoff, "The impact of a solid body which moves at the surface of an incompressible fluid," TCAHI Report No. 187, Moscow (1934).

surface, is struck by a body S which travels at a velocity $\bar{V}$, at an angle α from the horizontal plane. The fluid in the vicinity of the impact is set into motion; a small amount rises above the horizontal plane and forms a splash; a larger quantity is disturbed below the free surface, which is no longer horizontal, to make room for the body. The force of gravity, the hydrostatic force, viscous drag forces, surface tension forces, and the hydrodynamic forces all act on the body. But if the velocity $\bar{V}$ is large, during the impact stage, the hydrodynamic forces alone are significant, and the other forces are neglected in all that follows.

The hydrodynamic force is equal to the rate of change of momentum of the fluid set into motion by the impact. Since the fluid is assumed to be non-viscous and the motion is started from rest, the flow is irrotational and the velocity field may be derived from a potential Φ. The force is therefore:

$$\bar{F}=\frac{\rho}{\bar{V}}\frac{d}{dt}\int_{S+S'+S''}\Phi\frac{\partial\Phi}{\partial n}dS. \tag{1.1}$$

The integration is carried out over the boundaries of the fluid.** The forces acting on the body are therefore computed by a quadrature if the flow potential is known. Since the impact occurs in a very short time interval, $\bar{V}$ is assumed constant.

The velocity potential Φ of an incompressible perfect fluid satisfies the Laplace equation:

$$\Delta\Phi=0 \tag{1.2}$$

and certain conditions along the boundary of the fluid. The disturbance due to the impact must disappear far from the point of impact. Thus, if r denotes the distance from the point of impact, one has

$$\lim_{r\to\infty}\nabla\Phi=\lim_{r\to\infty}(\partial\Phi/\partial t)=0 \tag{1.3}$$

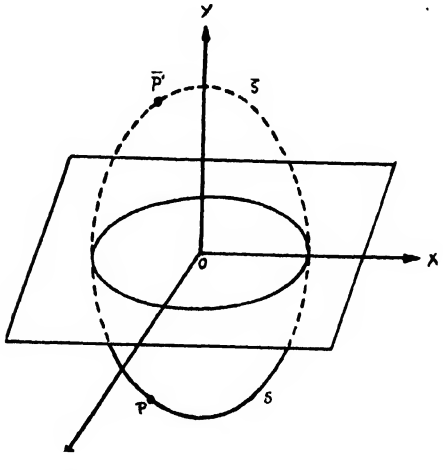


FIG. 1. General boundary conditions.

** S represents the surface of the submerged portion of the body, S' the free water surface, and S'' is a hemisphere of infinitely large radius with center at the body.

or the equivalent condition

$$\lim_{r\to\infty}\Phi=0. \tag{1.3a}$$

The velocity component normal to the surface of the impinging body must vanish:

$$\left(\frac{\partial\Phi}{\partial n}\right)_S=-V\cos\lambda, \tag{1.4}$$

where n is the direction normal to the wetted surface S, and λ is the angle between that normal and the direction of $\bar{V}$. Finally, along the free surface, the pressure in the fluid must equal atmospheric pressure, which may be taken equal to zero. This gives a condition on Φ, when Euler's equation is used:

$$(\partial\bar{u}/\partial t)+\bar{u}\cdot\nabla\bar{u}=(-1/\rho)\nabla p, \tag{1.5}$$

where $\bar{u}$ is the velocity vector in the fluid and p the pressure. An integration brings this equation into the form

$$-\partial\Phi/\partial t+\tfrac{1}{2}(\nabla\Phi)^2=-(p/\rho)+F(t). \tag{1.5a}$$

Boundary condition (1.3) applied on the free surface where p vanishes, gives $F(t)\equiv 0$, so that the boundary condition valid on the free surface is:

$$\tfrac{1}{2}(\nabla\Phi)^2-(\partial\Phi/\partial t)=0. \tag{1.5b}$$

One is thus led to investigate solutions of (1.2) subject to boundary conditions (1.3a, 1.4, 1.5b). Since (1.5b) is not linear, and the surface where it holds must be determined from the solution, this problem is extremely difficult; indeed, it has not been solved.

In linearizing boundary condition (1.5b), one argues that the impact occurs during a very short time interval, and may therefore be considered as an impulse. But it is known[10] that in the case of impulsive motion, the dynamic pressure is proportional to the velocity potential. Since the pressure vanishes on the free surface, one replaces condition (1.5b) by the more restrictive condition

$$\Phi(S')=0, \tag{1.5c}$$

where S' denotes the free surface. It is further assumed that the effect of the deformation of the free surface due to the splash is small, so that (1.5c) is applied along the original free surface:

$$\Phi(x,0,z,t)=0 \tag{1.5d}$$

in a cartesian system with origin at the center of impact, where the y axis is normal to the undisturbed free surface and the vector $\bar{V}$ lies in the plane $z=0$.

The problem of finding the potential Φ which satisfies Eq. (1.2) with boundary conditions (1.3a, 1.4, 1.5d) is a linear harmonic problem. It is convenient to investigate the potentials due to the various components

[10] H. Lamb, *Hydrodynamics* (Dover Publications, New York, 1944), sixth edition, London, (1932).

of $\bar{V}$ separately and to superimpose the results. Thus:

$$\Phi = V_x\varphi_1 + V_y\varphi_2 + V_z\varphi_3 + \Omega_{yz}\varphi_4 + \Omega_{zx}\varphi_5 + \Omega_{xy}\varphi_6 \quad (1.6)$$

where $V_{x,y,z}$ represent the x, y, z components of the velocity vector $\bar{V}$ and Ω_{ij} represent rotation velocities about the axes. In the problems to follow, V_z, Ω_{ij} vanish identically, so that

$$\Phi = V(\varphi_1 \cos\alpha + \varphi_2 \sin\alpha). \quad (1.6a)$$

If the symmetric image $\bar{S}$ of the wetted surface S with respect to the plane $y=0$ is constructed, (Fig. 1) and the field of flow is continued into the upper half-space, it follows from condition (1.5d) and the harmonic character of Φ that:

$$\Phi(x, y, z, t) = -\Phi(x, -y, z, t). \quad (1.5e)$$

The velocity components u, v, w, parallel to the x, y, z, axes, therefore have the following symmetry properties:

$$u(x, y, z) = -u(x, -y, z) \quad (1.7a)$$

$$v(x, y, z) = v(x, -y, z) \quad (1.7b)$$

$$w(x, y, z) = -w(x, -y, z) \quad (1.7c)$$

or, if the velocities normal to the surface $S+\bar{S}$ at two symmetric points $P, \bar{P}$ are compared,

$$(\partial\varphi/\partial n)_P = -(\partial\varphi/\partial n)_{\bar{P}}. \quad (1.7d)$$

If λ_1, λ_2 denote the angle between the normal to the surface $S+\bar{S}$ and the x, y axes, it follows from (1.7d) that boundary conditions (1.5, 1.4) become, for the potentials φ_1, φ_2

$$-\partial\varphi_1/\partial n = \cos\lambda_1 (|y|/y) \quad (1.8a)$$

$$-\partial\varphi_2/\partial n = \cos\lambda_2 \quad (1.8b)$$

these conditions being valid in the entire space.

The problem is thus split into two parts. The velocity component normal to the horizontal free surface induces the same potential as that of a fully submerged body in translation. The velocity component parallel to the horizontal free surface induces the potential which would obtain if the fluid above the plane of symmetry were flowing with a velocity equal and opposite that of the fluid below, causing a vortex sheet along the plane of symmetry.

One is now ready to express the momentum integral in terms of φ_1 and φ_2:

$$\frac{\rho}{V}\int_S \Phi \frac{\partial\Phi}{\partial n} dS' = \rho V \Big[\mathbf{i} \cos^2\alpha \int_S \varphi_1 \frac{\partial\varphi_1}{\partial n} dS + \sin\alpha\cos\alpha \Big(\mathbf{i}\int_S \varphi_2 \frac{\partial\varphi_1}{\partial n} dS + \mathbf{j}\int_S \varphi_1 \frac{\partial\varphi_2}{\partial n} dS \Big) + \mathbf{j}\sin^2\alpha \int_S \varphi_2 \frac{\partial\varphi_2}{\partial n} dS \Big] = \mathbf{i}M_x + (\mathbf{i}+\mathbf{j})M_{xy} + \mathbf{j}M_y, \quad (1.9)$$

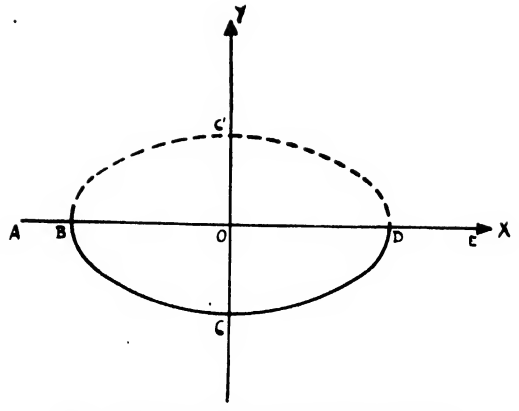


FIG. 2. Water impact of an elliptic cylinder.

where $\mathbf{i}$ and $\mathbf{j}$ are unit vectors parallel to the x, y axes. Note that φ_2 is positive over the entire surface S. On the other hand, φ_1 may be negative over part of the rear-half of the surface. If S is symmetric with respect to the plane $x=0$, φ_1 is antisymmetric and negative over the entire rear-half. But the dynamic pressure $p=\rho\Phi$ cannot become negative, since this would imply that the fluid is in tension and therefore separates from the body surface to form a cavity. Actually, because of the action of viscous forces and surface tension forces, separation is not instantaneous, and during the short period of time in which the impact occurs, the streamline pattern is not changed appreciably. But the integration in (1.9) must be carried out only over regions where $\varphi = \varphi_1\cos\alpha + \varphi_2\sin\alpha$ is positive.

The simplified problem formulated above does not depend on the time parameter directly since the equation of motion (1.2) and its boundary condition (1.3, 1.8) do not involve time variations. But as time passes, an increasing volume of the striking body is immersed in the fluid. Passage of time is, therefore, marked by a change in the dimensions and shape of the immersed body. Translation of the body in a horizontal direction does not change the shape of the immersed body and induces no force, but translation perpendicular to the free surface does.

In order to compute the time rate of change of momentum, therefore, it is necessary to determine how the immersed portion of the body changes with time. Most bodies whose impact may be of interest (projectiles, seaplane floats, etc.) have immersed portions of shapes so complicated that an analytic determination of φ_1, φ_2 is not possible. Following the scheme of previous calculations, one is led to approximate the immersed portion of the striking body at any time by some geometrically simple body (hemisphere, half-ellipsoid, etc.). The simple shape is defined by a certain number of parameters: the sphere by its radius, the ellipsoid of revolution by two axes, the general ellipsoid by three axes, etc. The approximating process consists in selecting certain properties of the actual body and defining the parameters of the approximating body so

that those properties are preserved. One can generally preserve as many properties as there are independent parameters defining the approximating body. For instance, if at any time t, the immersed portion of the striking body is approximated by a half-ellipsoid, the three axes of the ellipsoid are chosen so that the depth of penetration, the length and the immersed volume of the half-ellipsoid equal those of the actual body. As time passes, the depth of penetration, length and immersed volume of the actual body change, and the three axes of the approximating half-ellipsoid change correspondingly. The three axes of the ellipsoid are, therefore, known functions of time, when the motion of the actual body is known, either from measurements or from calculations. The momentum integral M (1.9a) of the approximating body depends only on its geometrical parameters $\alpha_1, \alpha_2, \alpha_3 \cdots \alpha_i$, which are known functions of time, since V and ρ are constant. The drag force is therefore given by:

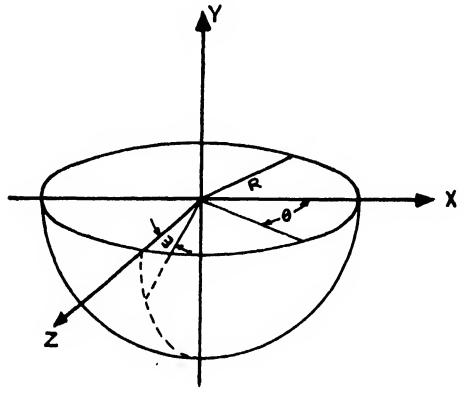


FIG. 3. Water impact of a sphere.

$$F_x=\rho V\cos\alpha\sum(d\alpha_i/dt)\times[(\partial M_x/\partial\alpha_i)\cos\alpha+(\partial M_{xy}/\partial\alpha_i)\sin\alpha] \quad (1.10a)$$

$$F_y=\rho V\sin\alpha\sum(d\alpha_i/dt)\times[(\partial M_{xy}/\partial\alpha_i)\cos\alpha+(\partial M_y/\partial\alpha_i)\sin\alpha]. \quad (1.10b)$$

The problem of finding approximately the forces acting on a given body when it strikes a horizontal surface is, therefore, investigated as follows:

An approximating body is selected and its parameters α_i are computed as functions of time from the known behavior of the actual body.

The potential problem for the approximating body is solved and the momentum integral is computed in terms of α_i.

The force function is found by application of Eqs. (1.10).

The second step of this procedure is the most important and difficult. It can be carried out once and for all for several approximating body shapes. Then the results can be applied to a wide variety of problems by the proper fitting procedure.

In the following, the potential problem is solved for an infinitely wide semi-elliptic cylinder (two-dimensional problem solved by Sedoff's methods) and for a sphere, an ellipsoid of revolution and a general ellipsoid (solved by harmonic analysis methods). As an example of the fitting process, the water entry forces on a sphere striking at 45° are computed by using a general ellipsoid as the approximating body.

2. INFINITE SEMI-ELLIPTIC CYLINDER

The first potential problem discussed in detail concerns an infinitely wide semi-elliptic cylinder of major axis $2a$ and minor axis $2b$, the free surface $y=0$ coinciding with the major axis (Fig. 2). The boundary conditions to be satisfied by the potentials φ_1, φ_2 in the entire x, y plane are obtained from (1.8). If a stream function ψ is introduced, then, on the body surface,

$$\partial\varphi/\partial n=\partial\psi/\partial s=\cos\alpha(\partial y/\partial s)(|y|/y)-\sin\alpha(\partial x/\partial s), \quad (2.1)$$

where ds is a length element along the surface of the ellipse. This is integrated to give:

$$\psi=|y|\cos\alpha-x\sin\alpha. \quad (2.1a)$$

One is now required to find an analytic complex potential function $F(z)=F(x+iy)$ whose imaginary part satisfies (2.1a) on the boundary of the ellipse.

The conformal transformation

$$-2z=(a-b)\zeta+(a+b)1/\zeta \quad (2.2)$$

transforms the ellipse in the z plane into the unit circle in the ζ-plane. Therefore, in the ζ-plane the complex potential $F(\zeta)$ must satisfy the conditions below on the unit circle $\zeta=e^{i\theta}$:

$$\psi=b|\sin\theta|\cos\alpha+a\cos\theta\sin\alpha=\frac{2b\cos\alpha}{\pi}\left[1-2\sum_1^\infty\frac{\cos 2n\theta}{4n^2-1}\right]+a\sin\alpha\cos\theta. \quad (2.3)$$

It follows that:

$$F(\zeta)=\frac{2bi\cos\alpha}{\pi}\left[1+2\sum_1^\infty\frac{\zeta^{2n}}{4n^2-1}\right]+ai\sin\alpha\zeta. \quad (2.4)$$

To sum the power series inside and on the unit circle, note that

$$2\sum_1^\infty\frac{\zeta^{2n}}{4n^2-1}=\zeta\sum_1^\infty\frac{\zeta^{2n-1}}{2n-1}-\frac{1}{\zeta}\sum_1^\infty\frac{\zeta^{2n+1}}{2n+1} \quad (2.5a)$$

and since

$$\sum_1^\infty\zeta^{2n}=\zeta^2/1-\zeta^2 \quad (2.5b)$$

one has

$$\sum_1^\infty\frac{\zeta^{2n-1}}{2n-1}=\int_0^\zeta\frac{d\zeta}{1-\zeta^2}=\tfrac{1}{2}\log\frac{1+\zeta}{1-\zeta} \quad (2.5c)$$

$$\sum_1^\infty\frac{\zeta^{2n+1}}{2n+1}=\int_0^\zeta\frac{\zeta^2d\zeta}{1-\zeta^2}=-\zeta+\tfrac{1}{2}\log\frac{1+\zeta}{1-\zeta} \quad (2.5d)$$

so that

$$F(\zeta)=\frac{ib\cos\alpha}{\pi}\left(\frac{1}{\zeta}-\zeta\right)\log\frac{1+\zeta}{1-\zeta}+ia\zeta\sin\alpha. \quad (2.6)$$

All the manipulations carried out above are valid in the annulus $0<|\zeta|<1$ because all series and integrals converge uniformly there. It is easily verified that $F(\zeta)$ is an analytic function regular everywhere in and on the unit circle. Separating real and imaginary parts, on the circle, one finds

$$\psi=b\cos\alpha|\sin\theta|+a\cos\theta\sin\alpha \quad (2.7a)$$

$$\varphi=-\sin\theta[(2b/\pi)\cos\alpha\log|\tan(\theta/2)|+a\sin\alpha]. \quad (2.7b)$$

The boundary condition (2.3) is therefore satisfied and the pressure distribution is given by (2.7b). The separation point occurs at θ_s defined by

$$(2b/\pi)\cos\alpha\log|\tan(\theta_s/2)|+a\sin\alpha=0 \quad (2.8a)$$

or

$$\theta_s=2\tan^{-1}\exp[-(a\pi/2b)\tan\alpha]. \quad (2.8b)$$

In most cases $a\gg b$ and $\tan\alpha=0(1)$; then θ_s is small: $\theta_s=0(10^{-3})$ and the separation phenomenon may be neglected. Under those circumstances, the momentum integrals are:

$$M_x=2b^2/\pi \quad (2.9a)$$

$$M_{xy}=M_{yz}=0 \quad (2.9b)$$

$$M_y=\pi a^2/2. \quad (2.9c)$$

3. THE SPHERE

Consider a sphere of radius R half-submerged in a fluid (Fig. 3). The potential $\varphi=\varphi_1\cos\alpha+\varphi_2\sin\alpha$ must satisfy the equation

$$\Delta\varphi_1=\Delta\varphi_2=0 \quad (3.1)$$

with the boundary conditions:

$$\lim_{r\to\infty}\varphi_1=\lim_{r\to\infty}\varphi_2=0 \quad (3.2)$$

$$(\partial\varphi_2/\partial r)(R,\theta,\omega)=-\cos\omega \quad (3.3a)$$

$$(\partial\varphi_1/\partial r)(R,\theta,\omega)=-(|\cos\omega|/\cos\omega)\cos\theta. \quad (3.3b)$$

The potential φ_2 is immediately found:

$$\varphi_2=(R^3/2r^2)\cos\omega. \quad (3.4)$$

The potential φ_1 is determined by fitting the general harmonic function convergent for large r

$$\varphi_2=\sum_0^\infty\sum_0^{n+1}\frac{1}{r^{n+1}}P_n^s(\cos\omega)[a_n^s\cos s\theta+b_n^s\sin s\theta] \quad (3.5)$$

to satisfy boundary condition (3.3b). It is clear by inspection that $b_n^s=0$ and $a_n^{s>1}=0$, while the non-vanishing coefficients a_n^1 must satisfy the relation:

$$\sum_0^\infty\frac{n+1}{R^{n+2}}a_n^1P_n^1(\cos\omega)=\sum_0^\infty\alpha_nP_n^1$$

$$=-1 \quad [0<|\omega|<(\pi/2)]$$

$$=1 \quad [(\pi/2)<|\omega|<\pi]. \quad (3.6)$$

If Eq. (3.6) is multiplied by $P_k^1(\cos\omega)\sin\omega$ and integrated from 0 to π, all terms in the series except the k

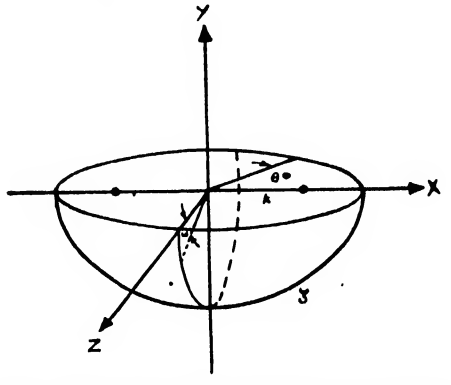


FIG. 4. Water impact of an ellipsoid of revolution.

term vanish because of the orthogonality of Legendre polynomials. One, therefore, obtains:

$$\alpha_k=\frac{2k+1}{2k(k+1)}\left[\int_0^{\pi/2}P_k^1(\cos\omega)\sin\omega d\omega-\int_{A/2}^{\pi}P_k^1(\cos\omega)\sin\omega d\omega\right]. \quad (3.7a)$$

Since $P_k^1(\cos\omega)=\sin\omega[dP_k(\cos\omega)/d(\cos\omega)]$, this is rewritten, with due regard to symmetry properties:

$$\alpha_{2k}=\frac{4k+1}{2k(2k+1)}\int_0^{\pi/2}\sin^2\omega P_{2k}'(\cos\omega)d\omega. \quad (3.7b)$$

$$\alpha_{2k+1}=0$$

By use of Rodrigues's Formula[11] the Legendre polynomial $P_{2k}(z)$ can be written as:

$$P_{2k}(z)=\frac{1}{2^kk!}\sum_{r=0}^{k}\binom{k-r}{r}\frac{(2k+2r-1)!!}{(2r-1)!!}(-)^{k-r}z^{2r}. \quad (3.8a)$$

It is also known[12] that

$$\int_0^{\pi/2}\sin^2\omega\cos^{2r-1}\omega d\omega=\frac{(2r-2)!!}{(2r+1)!!}. \quad (3.8b)$$

[11] Whittaker and Watson, *Modern Analysis* (Cambridge University Press, London, 1940), fourth edition.

[12] H. B. Dwight, *Tables of Integrals and Other Mathematical Data* (MacMillan Company, Ltd., London, 1934).

Combining formulas (3.8a) and (3.8b), one obtains:

$$\alpha_{2k}=\frac{4k+1}{2^k k!(2k+1)}\sum_{r=1}^{k}(-)^{k-r+1}\times\binom{k-r}{k}\frac{r(2r-2)!!(2k+2r-1)!!}{(2r-1)!!(2r+1)!!}. \quad (3.9)$$

Thus, for instance:

$$\begin{aligned}&\alpha_2=5/6=0.833 && \alpha_8=-0.041\\ &\alpha_4=-(3/40)=-0.075 && \alpha_{10}=0.0658.\\ &\alpha_6=0.147 && \alpha_{12}=-0.0094\cdots\end{aligned} \quad (3.9a)$$

The convergence of the series (3.6) is established by noting that the integral in (3.7b) can be determined asymptotically as $k\to\infty$ as follows:

$$\begin{aligned}\lim_{k\to\infty}\int_0^{\pi/2} P_{2k}'(\cos\omega)\sin^2\omega d\omega &= \int_0^{\pi/2}\sin^2\omega\lim_{k\to\infty}P_{2k}'(\cos\omega)d\omega\\ &= \int_0^{\pi/2}\sin^2\omega(k/\pi)^{\frac{1}{2}}\sin 2k\omega d\omega\\ &= \tfrac{1}{4}(k/\pi)^{\frac{1}{2}}\int_0^{\pi}(1-\cos\theta)\sin k\theta d\theta\sim 1/(k)^{\frac{1}{2}}. \end{aligned} \quad (3.10)$$

The general coefficient α_k therefore behaves as $k^{-\frac{1}{2}}$. Since[13]

$$\lim_{k\to\infty}P_{2k}^{1}(\cos\theta) = -2(k/\pi\sin\theta)^{\frac{1}{2}}\sin[(2k+\tfrac{1}{2})\theta+3\pi/4] \quad (3.10')$$

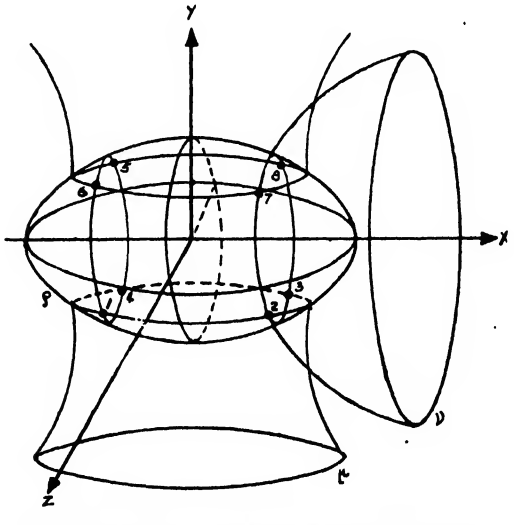


Fig. 5. Coordinate system for a general ellipsoid.

[13] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics* (Cambridge University Press, London, 1931), Chapter 11.

the series (3.6) behaves as $\sum(e^{ik\theta}/k)$ which is known to converge.

The potential of the horizontal flow past a sphere of radius R is, therefore,

$$\varphi_1=R\cos\theta\sum_1^{\infty}(R/r)^{n+1}(\alpha_n/n+1)P_n^1(\cos\omega). \quad (3.11)$$

The separation point is reached along the curve $\theta(\omega)$ where the following relation holds:

$$\theta=\sec^{-1}[-2\cot\alpha\sec\omega\sum_1^{\infty}(\alpha_n/n+1)P_n^1(\cos\omega)]. \quad (3.12)$$

If separation effects are neglected, one finds the following momentum integrals:

$$M_x=(\pi/2)\rho R^3\sum_0^{\infty}(\alpha_n)^2 \quad (3.13a)$$

$$M_{xy}=M_{yx}=0 \quad (3.13b)$$

$$M_y=(\pi/3)\rho R^3. \quad (3.13c)$$

4. HALF AN ELLIPSOID OF REVOLUTION

A slight generalization is obtained if the hemisphere considered above is replaced by an ellipsoid of revolution whose axis is parallel to the horizontal velocity component (Fig. 4). The potential problem is most simply discussed in terms of elliptic coordinates. If x, y, z, are cartesian coordinates as before, and the distance between the foci of the ellipsoid is $2k$, the new coordinates μ, ζ, ω are defined by

$$x=k\mu\zeta \quad (4.1a)$$

$$y=k[(1-\mu^2)(\zeta^2-1)]^{\frac{1}{2}}\sin\omega \quad (4.1b)$$

$$z=k[(1-\mu^2)(\zeta^2-1)]^{\frac{1}{2}}\cos\omega. \quad (4.1c)$$

The ellipsoid under study is defined by a value ζ_0 of the parameter ζ, and solutions are sought in the domain $\zeta>\zeta_0$ which corresponds to the space outside the ellipsoid.

The boundary conditions corresponding to (1.8a, b) are here:

$$\left(\frac{\partial\varphi_1}{\partial\zeta}\right)_{\zeta_0}=-k\mu\frac{|\cos\omega|}{\cos\omega} \quad (4.2a)$$

$$\left(\frac{\partial\varphi_2}{\partial\zeta}\right)_{\zeta_0}=-k\zeta_0\left(\frac{1-\mu^2}{\zeta_0^2-1}\right)^{\frac{1}{2}}\sin\omega. \quad (4.2b)$$

The potential φ_2 is known to be[11]

$$\varphi_2=\frac{k}{\tfrac{1}{2}\log(\zeta_0+1/\zeta_0-1)-[\zeta_0^2-2/\zeta_0(\zeta_0^2-1)]}\times\left\{\tfrac{1}{2}\log\frac{\zeta+1}{\zeta-1}-\frac{\zeta}{\zeta^2-1}\right\}[(1-\mu^2)(\zeta^2-1)]^{\frac{1}{2}}\cos\omega. \quad (4.3)$$

The general potential convergent for $\zeta > \zeta_0$ is

$$\varphi_1 = \sum_0^\infty \sum_1^{n+1} Q_n^s(\zeta) P_n^s(\mu)[a_n^s \sin s\omega + b_n^s \cos s\omega] \quad (4.4)$$

so that boundary condition (4.2a) gives:

$$\sum_0^\infty \sum_1^{n+1} Q_n'^s(\zeta_0) P_n^s(\mu) a_n^s \sin s\omega$$

$$= \sum_0^\infty \sum_1^{n+1} \alpha_n^s P_n^s(\mu) \sin s\omega = -\mu(-\pi < \omega < 0) \quad (4.5a)$$

$$= \mu(0 < \omega < \pi)$$

from which the coefficients α_n^s are determined. Multiplication of (4.5a) by $\sin(r\omega)$ and integration between $-\pi$ and π leads to:

$$\sum_0^\infty \alpha_n^{2r-1} P_n^{2r-1}(\mu) = -\frac{4}{\pi} \frac{(-)^r \mu}{2r+1} \quad (4.5b)$$

$$\alpha_n^{2r} = 0.$$

Multiplication of (4.5b) by $P_k^{2r+1}(\mu)$ and integration between -1 and 1 leads to:

$$\alpha_{2k}^{2r-1} = \frac{(-)^r 8(4k+1)(2k-2r-1)!}{\pi(2r+1)(2k+2r+1)!} \times \int_0^1 \mu P_{2k}^{2r+1}(\mu) d\mu \quad (4.5c)$$

$$\alpha_{2k+1}^{2r+1} = 0.$$

After two integrations by parts, the integral in (4.5c) becomes:

$$\int_0^1 \mu P_{2k}^{2r+1}(\mu) d\mu = \int_0^1 \mu(1-\mu^2)^{\frac{1}{2}} \frac{d}{d\mu} P_{2k}^{2r}(\mu) d\mu$$

$$= \int_0^1 \frac{2\mu^2 - 1}{(1-\mu^2)^{\frac{1}{2}}} P_{2k}^{2r}(\mu) d\mu$$

$$= \int_0^1 (2k^2 - 1) \frac{d}{d\mu} P_{2k}^{2r-1}(\mu) d\mu$$

$$= -4 \int_0^1 \mu P_{2k}^{2r-1}(\mu) d\mu \quad (4.6a)$$

since

$$P_{2k}^{2r+1}(0) = P_{2k}^{2r+1}(1) = 0.$$

The recursion formula (4.6a) gives at once:

$$\int_0^1 \mu P_{2k}^{2r+1}(\mu) d\mu = (-4)^r \int_0^1 \mu P_{2k}^1(\mu) d\mu$$

$$= (-4)^r \int_0^{\pi/2} \sin^2\theta P_{2k}'(\cos\theta) d\theta, \quad (4.6b)$$

By using the representation (3.8a) for $P_{2k}(z)$ and the relation:

$$\int_0^{\pi/2} \sin^2\theta \cos^{2p}\theta d\theta = \frac{\pi}{2} \frac{(2p-1)!!}{(2p+2)!!} \quad (4.6c)$$

one finds the coefficient α_{2k}^{2r+1}

$$\alpha_{2k}^{2r+1} = \frac{4^{r+1}}{2^{k-1}k!} \frac{4k+1}{2r+1} \frac{(2k-2r-1)!}{(2k+2r+1)!} \times \sum_{p=1}^k (-)^{k-p+1} \binom{k-p}{p} \frac{p(2k+2p-1)!!}{(2p+2)!!} \quad (4.7)$$

Thus for instance:

$$\begin{array}{lll} \alpha_2^1 = 1.250 & & \\ \alpha_4^1 = 0.281 & & \\ \alpha_6^1 = 0.127 & \alpha_4^3 = 0.00161 & \alpha_6^5 = 0.00001. \\ \alpha_8^1 = 0.0726 & \alpha_6^3 = 0.00002 & \\ \alpha_{10}^1 = 0.0471 & & \\ \alpha_{12}^1 = 0.0328 & & \end{array} \quad (4.7a)$$

The uniform convergence of the doubly infinite series (4.5a) is established by the Weierstrass M test if $\sum\sum \alpha_k^r$ converges: The integral in (4.5c) converges:

$$\lim_{k \to \infty} \left| \int_0^1 \mu P_{2k}^1 d\mu \right| < \frac{A}{(k)^{\frac{1}{2}}} \quad (4.8a)$$

as can be shown by an argument similar to that of (3.10). For any value of r, therefore, the series

$$\sum_{k=0}^\infty \alpha_k^r = S'(r)$$

converges absolutely. For any given value of k,

$$\lim_{r \to \infty} \sum_0^k \alpha_k^r < A(k) \sum \frac{4^r}{(2r+1)(4r)!} \quad (4.8b)$$

so that the sum of rows as well as the sum of columns converges absolutely. Then, one has

$$|\sum\sum \alpha_k^r| < \sum\sum (1/(k+r)!) \quad (4.8c)$$

which establishes the convergence of the series. The potential φ_1 for the horizontal flow past the half-ellipsoid of revolution is, therefore:

$$\varphi_1 = \sum_{n=1}^\infty \sum_{s=0}^{n-1} \frac{Q_{2n}^{2s+1}(\zeta)}{Q_{2n}'^{2s+1}(\zeta_0)} \times \alpha_{2n}^{2s+1} P_{2n}^{2s+1}(\mu) \sin(2s+1)\omega. \quad (4.9)$$

The momentum integrals are:

$$M_x = -\frac{\pi}{4} k^3(\zeta_0^2 - 1) \sum_{n=1}^\infty \sum_{s=1}^{n-1} \frac{Q_{2n}^{2s+1}(\zeta_0)}{Q_{2n}'^{2s+1}(\zeta_0)} \times \frac{(2n+2s+1)!}{(2n-2s-1)!} \frac{(\alpha_{2n}^{2s+1})^2}{4n+1} \quad (4.10a)$$

$$M_{xy}=M_{yz}=0 \tag{4.10b}$$

$$M_y=\frac{\pi k^3}{2}\zeta_0(\zeta_0^2-1)$$

$$\times\frac{\log(\zeta_0+1/\zeta_0-1)-(2\zeta_0/\zeta_0^2-1)}{\log(\zeta_0+1/\zeta_0-1)-[2(\zeta_0^2-2)/\zeta_0(\zeta_0^2-1)]}. \tag{4.10c}$$

Note that the entire argument was carried through for a prolate ellipsoid ($\zeta>1$). If the ellipsoid is oblate ($\zeta<1$), ζ is replaced by $i\zeta$ and the function $q(\zeta)$ replaces $Q(\zeta)$, where

$$g_n^s(\zeta)=Q_n^s(i\zeta) \tag{4.11}$$

everything else remaining unchanged.

5. HALF A GENERAL ELLIPSOID

A further generalization is obtained by considering an ellipsoid with focal length $2k$ parallel to the horizontal velocity component, and focal length $2h$ horizontal and normal to it. The problem is now formulated in terms of ellipsoidal coordinates defined by E. W. Hobson:[13]

$$x=\rho\mu\nu/hk \tag{5.1a}$$

$$y=\frac{[(\rho^2-h^2)(\mu^2-h^2)(h^2-\nu^2)]^{\frac{1}{2}}}{h(k^2-h^2)^{\frac{1}{2}}} \tag{5.1b}$$

$$z=\frac{[(\rho^2-k^2)(k^2-\mu^2)(k^2-\nu^2)]^{\frac{1}{2}}}{k(k^2-h^2)^{\frac{1}{2}}}. \tag{5.1c}$$

The surfaces $\rho=$const. represent confocal ellipsoids; ρ_0 in particular represents the ellipsoid under investigation; $\mu=$const. are hyperboloids of one sheet and $\nu=$const. are hyperboloids of two sheets. The coordinate system is shown on Fig. 5. Any point is defined uniquely if the signs of the radicals in (5.1) are defined. Boundary conditions (1.8a, b) become in this system:

$$\left(\frac{\partial\varphi_1}{\partial\rho}\right)_{\rho_0}=-\frac{\mu\nu}{hk}\frac{|y|}{y} \tag{5.2a}$$

$$\left(\frac{\partial\varphi_2}{\partial\rho}\right)_{\rho_0}=-\frac{\rho_0}{h}\left\{\frac{[(\mu^2-h^2)(h^2-\nu^2)]}{[(\rho_0^2-h^2)(k^2-h^2)]}\right\}^{\frac{1}{2}}. \tag{5.2b}$$

The potential φ_2 is given explicitly in the standard literature[11] by the formula:

$$\varphi_2=\frac{\rho_0}{2h}\left[\frac{(\rho_0^2-h^2)(\rho_0^2-k^2)}{k^2-h^2}\right]^{\frac{1}{2}}$$

$$\times\frac{1}{1-\rho_0[(\rho_0^2-h^2)(\rho_0^2-k^2)]^{\frac{1}{2}}\int_{\rho_0}^{\infty}\frac{d\rho}{(\rho^2-h^2)^{\frac{1}{2}}(\rho^2-k^2)^{\frac{1}{2}}}}$$

$$\times[(\rho^2-h^2)(\mu^2-h^2)(h^2-\nu^2)]^{\frac{1}{2}}$$

$$\times\int_{\rho}^{\infty}\frac{d\rho}{(\rho^2-h^2)^{\frac{1}{2}}(\rho^2-k^2)^{\frac{1}{2}}}. \tag{5.3}$$

The general potential convergent for $\rho>\rho_0$ is

$$\varphi_1=\sum_0^\infty\sum_1^{2n+1}a_n^sF_n^s(\rho)E_n^s(\mu)E_n^s(\nu), \tag{5.4a}$$

where F_n^s, E_n^s are general Lamé functions, convergent as $\rho\to\infty$ and μ, $\nu\to0$, and satisfying Lamé's differential equation:

$$(\lambda^2-h^2)(\lambda^2-k^2)(d^2E_n/d\lambda^2)$$
$$+\lambda(2\lambda^2-h^2-k^2)(dE_n/d\lambda)$$
$$+[(h^2+k^2)p-n(n+1)\lambda^2]E_n=0 \tag{5.4b}$$
$$\lambda=\rho,\ \mu,\ \nu$$

the parameter n is an integer, while $p(h, k, n)$ is chosen so that E_n^s is a polynomial. The theory of Lamé functions is discussed in detail in.[13]

Boundary condition (5.2a) gives here:

$$hk\sum_0^\infty\sum_1^{2n+1}F_n'^s(\rho_0)a_n^sE_n^s(\mu)E_n^s(\nu)$$

$$=\sum_0^\infty\sum_1^{2n+1}\alpha_n^sE_n^s(\mu)E_n^s(\nu)=\mu\nu[(\mu^2-h^2)^{\frac{1}{2}}>0]$$

$$=-\mu\nu[(\mu^2-h^2)^{\frac{1}{2}}<0]. \tag{5.5}$$

From the form of the function to be expanded, it is clear that only odd Lamé functions K_{2n+1}^s of the form

$$K_{2n+1}^s(\lambda)=\beta_1^s\lambda+\beta_3^s\lambda^3+\cdots\beta_{2n+1}^s\lambda^{2n+1} \tag{5.6a}$$

will appear in φ_1. The coefficients β_n^s are defined by recurrence formulas and depend on (h, k, p). They have been tabulated by Guerritore.[14] It is then known[15] that the coefficients α_n^s are given by:

$$\alpha_{2n+1}^s=\frac{\int_0^h\int_h^k(\mu^2-\nu^2)\mu\nu K_{2n+1}^s(\mu)K_{2n+1}^s(\nu)\frac{d\mu}{[(h^2-\mu^2)(\mu^2-k^2)]^{\frac{1}{2}}}\frac{d\nu}{[(h^2-\nu^2)(k^2-\nu^2)]^{\frac{1}{2}}}}{\int_0^h\int_h^k(\mu^2-\nu^2)[K_{2n+1}^s(\mu)K_{2n+1}^s(\nu)]^2\frac{d\mu}{[(h^2-\mu^2)(\mu^2-k^2)]^{\frac{1}{2}}}\frac{d\nu}{[(h^2-\nu^2)(k^2-\nu^2)]^{\frac{1}{2}}}}. \tag{5.7}$$

[14] M. Guerritore, Table of Lamé functions, Giornale de Matematica (2) XVI, 164–172 (1909).

[15] O. Volk, "Uber die entwicklung von functionen zweier complexen veranderlichen nach lameschen functionen," Math. Zeitsch 23, 224 (1925).

If the expansion (5.6) for $K_{2n+1}{}^{s}(\lambda)$ is substituted into the integrals (5.7), it is found that one must investigate integrals of the form:

$$A_{2r}(\lambda)=\int_0^h \frac{\lambda^{2r}d\lambda}{[(h^2-\lambda^2)(k^2-\lambda^2)]^{\frac{1}{2}}} \tag{5.8a}$$

$$B_{2r}(\lambda)=\int_h^k \frac{\lambda^{2r}d\lambda}{[(\lambda^2-h^2)(k^2-\lambda^2)]^{\frac{1}{2}}}. \tag{5.8b}$$

It is easily verified that:

$$A_0=(1/k)F(h/k) \qquad B_0=(1/k)F\left[\frac{(k^2-h^2)^{\frac{1}{2}}}{k}\right] \tag{5.9a,b}$$

$$A_2=k(F(h/k)-E(h/k)) \quad B_2=kE\left[\frac{(k^2-h^2)^{\frac{1}{2}}}{k}\right]. \tag{5.9c,d}$$

Repeated integrations by parts then lead to the recurrence formulas:

$$(2n-1)A_{2n}=(2n-4)(h^2+k^2)A_{2n-2} -(2n-3)h^2k^2A_{2n+4} \tag{5.10}$$

and an identical result for B_{2n}. The integrals A_{2n}, B_{2n} are, therefore, expressed as functions of the parameters h, k in terms of elliptic integrals.

The numerator and denominator of (5.7) are now simply expressed in terms of the known coefficients $\beta_n{}^{s}$ and the integrals A_{2n}, B_{2n}.

$$N_{2n+1}{}^{s}=(\beta_1{}^{s})^2(B_4A_2-B_2A_4) +\beta_1{}^{s}\beta_3{}^{s}(B_6A_2-A_6B_2) +(\beta_3{}^{s})^2(B_6A_4-B_4A_6)\cdots \tag{5.11a}$$

$$D_{2n+1}{}^{s}=(\beta_1{}^{s})^4(B_4A_2-B_2A_4) +2\beta_3{}^{s}(\beta_1{}^{s})^3(B_6A_2-A_6B_2)\cdots \tag{5.11b}$$

$$\alpha_{2n+1}{}^{s}=N_{2n+1}{}^{s}/D_{2n+1}{}^{s}. \tag{5.11c}$$

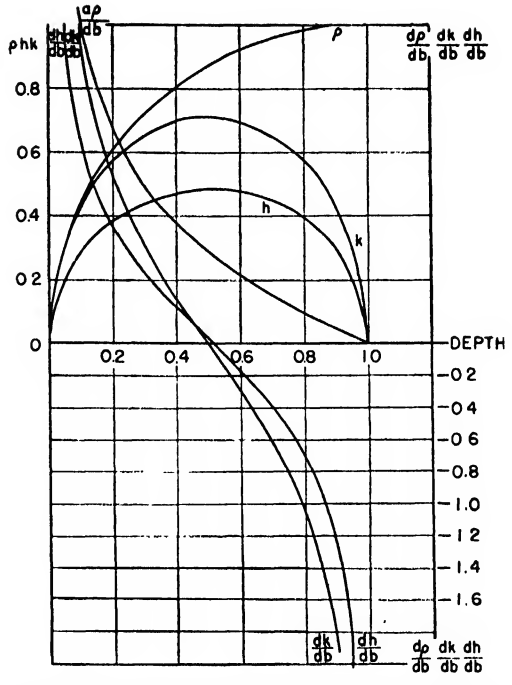


FIG. 6. The approximating functions for a partially submerged sphere and their derivatives.

The potential function φ_1 is then given by:

$$\varphi_1=\frac{1}{hk}\sum_0^{\infty}\sum_1^{n+1/2}\frac{F_{2n+1}{}^{s}(\rho)}{F_{2n+1}{}'^{s}(\rho_0)} \times\alpha_{2n+1}{}^{s}K_{2n+1}{}^{s}(\mu)K_{2n+1}{}^{s}(\nu). \tag{5.12}$$

The momentum integrals are:

$$M_x=\frac{[(\rho_0^2-h^2)(\rho_0^2-k^2)]^{\frac{1}{2}}}{h^2k^2}\sum_0^{\infty}\sum_1^{n+1/2}\frac{F_{2n+1}{}^{s}(\rho_0)}{F_{2n+1}{}'^{s}(\rho_0)}(\alpha_{2n+1}{}^{s})^2D_{2n+1}{}^{s} \tag{5.13a}$$

$$M_{xy}=M_{yz}=0 \tag{5.13b}$$

$$M_y=(\pi\rho_0/12)[(\rho_0^2-h^2)(\rho_0^2-k^2)]^{\frac{1}{2}}\frac{\rho_0[(\rho_0^2-h^2)(\rho_0^2-k^2)]^{\frac{1}{2}}\int_{\rho_0}^{\infty}\frac{d\rho}{\rho_0(\rho^2-h^2)^{\frac{1}{2}}(\rho^2-k^2)^{\frac{1}{2}}}}{1-\rho_0[(\rho_0^2-h^2)(\rho_0^2-k^2)]^{\frac{1}{2}}\int_{\rho_0}^{\infty}\frac{d\rho}{\rho_0(\rho^2-h^2)^{\frac{1}{2}}(\rho^2-k^2)^{\frac{1}{2}}}}. \tag{5.13c}$$

The first three terms of this expansion were computed numerically and used in Section 6.

6. WATER IMPACT OF A SPHERE AT AN ENTRY ANGLE OF 45°

The results of the previous pages are now applied to a specific problem. The drag on a sphere of radius R which strikes the water surface at 45° is computed as a function of time. The submerged portion of the sphere is approximated by a half-ellipsoid in such a manner that the depth of penetration d of the sphere equals the minor axis, the diameter of the circle of intersection between the sphere and the free surface equals the major axis and the submerged volume of the sphere, V, equal the volume of the semi-ellipsoid. If the dimensionless length ξ is introduced, one has:

$$\xi=tV\sin\alpha/R \tag{6.1a}$$

$$\rho_0(t)/R=[\xi(2-\xi)]^{\frac{1}{2}} \tag{6.1b}$$

$$k(t)/R=[2\xi(1-\xi)]^{\frac{1}{2}} \tag{6.1c}$$

$$h(t)/R=\tfrac{1}{2}\left[\frac{\xi(1-\xi)(7-3\xi)}{2-\xi}\right]^{\frac{1}{2}} \tag{6.1d}$$

$$d\alpha_i/dt=(V\sin\alpha/R)(d\alpha_i/d\xi). \tag{6.1e}$$

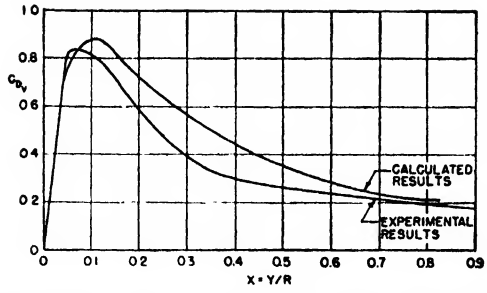


FIG. 7. Vertical component of drag of a sphere of unit radius in water impact $C_D = F_Y/\pi R^2\rho(V^2/2)$.

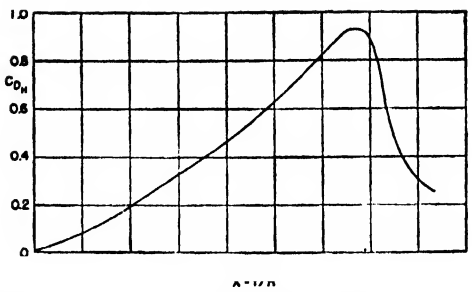


FIG. 8. Horizontal component of drag of a sphere in water impact.

The force functions are given by formulas (1.10a, b) and the derivatives are given explicitly as:

$$d\rho_0/dt = V\sin\alpha\frac{1-\xi}{2[\xi(2-\xi)]^{\frac{1}{2}}} \tag{6.2a}$$

$$dk/dt = V\sin\alpha\frac{1-2\xi}{[2\xi(1-\xi)]^{\frac{1}{2}}} \tag{6.2b}$$

$$dh/dt = V\sin\alpha\frac{7-20\xi+14\xi^2-3\xi^3}{2[\xi(1-\xi)(7-3\xi)(2-\xi)^2]^{\frac{1}{2}}} \tag{6.2c}$$

while the derivatives $\partial M/\partial\alpha_i$ are obtained by differentiating (5.13) with respect to ρ_0, h, k. This formulation is valid only while $\xi<1$, or up to the time when the sphere is half-submerged. After that instant, flow past an ellipsoid does not give a good approximation to the streamline pattern, and a cavity is beginning to develop.

The results of the calculation for the sphere are shown on Figs. 6, 7, 8. It is found that the value of the vertical drag component agrees fairly well with the value obtained in unpublished experiments carried out at the Morris Dam Hydrodynamics Station; it also gives the same type of result as that obtained by Schiffman and Spencer. The horizontal component of drag reaches its greatest value at a later point.

CONCLUSION

The impact of a body of arbitrary shape at an arbitrary angle was investigated by approximate methods. The two main approximations consisted in linearizing the boundary condition at the free surface by neglecting the splash and in replacing the submerged portion of the body at any time by an ellipsoid of identical depth of penetration, length, and submerged volume. As an illustration of the method, the impact of a sphere at a 45° angle was studied in some detail. The vertical force component turned out to be in good agreement with experimental results. The behavior of the horizontal component, which has not been measured, was also computed and shown.

Depolymerization by Ultrasonic Irradiation: The Role of Cavitation

ALFRED WEISSLER
Naval Research Laboratory, Washington, D. C.
(Received July 7, 1949)

Contrary to current belief, cavitation has been found responsible for the depolymerizing effect of intense ultrasonic waves. This was demonstrated by irradiating two portions of a 1 percent polystyrene solution in toluene under conditions identical except for the following. The first portion was given no special prior treatment, showed many cavitation bubbles during irradiation, and decreased in molecular weight (as measured by the intrinsic viscosity) to about one-tenth of the initial value. The second portion was given a preliminary treatment of degassing by boiling under vacuum, showed no cavitation bubbles during the irradiation, and underwent no appreciable change in molecular weight.

Similar experiments with solutions of hydroxyethyl cellulose in water showed that, in this case also, cavitation is necessary for depolymerization. The opposite conclusion of earlier investigators is attributed to their inadequate method for eliminating cavitation.

Oxidants known to be produced by ultrasonic waves in solutions containing dissolved oxygen or nitrogen cannot be responsible for the degradation, because substantially the same amount of depolymerization occurs even when helium is the only gas present.

INTRODUCTION

SOLUTIONS of high polymers may be affected by ultrasonic irradiation in two different ways.[1,2] First, Freundlich and Gillings found[3] that weak irradiation of fresh solutions of gelatin in water causes a considerable reduction in the structural or non-Newtonian viscosity. But on standing for several days, the solution regains almost as high a viscosity as it had before irradiation. This effect is believed due to the temporary breaking of the loose gel network of van der Waals bonds between adjacent polymer molecules. It does not take place to an appreciable extent in the absence of cavitation.[4]

The second way in which ultrasonic waves may affect high polymers is to cause depolymerization—the actual breaking of chemical bonds in the chain, to give molecules of smaller size than originally, and therefore a *permanent* decrease in viscosity. This was studied in detail by Schmid and Rommel[5] in order to determine whether cavitation is responsible for this effect also. In a typical experiment, they found that a solution of nitrocellulose in butyl acetate underwent considerable degradation when irradiated in a steel bomb for an hour under atmospheric pressure, but less degradation under 5 atmos. of nitrogen, and none at all under 8 atmos. of oxygen. Similarly, a solution of polystyrene in toluene suffered much depolymerization when irradiated under atmospheric pressure, less under 8 atmos. of O_2, and still less under 15 atmos. of O_2. Schmid concluded that depolymerization occurs in the complete absence of cavitation, because as little as 5 atmos. excess pressure would have suppressed cavitation caused by the moderate-intensity ultrasonic waves used.

However, this conclusion appears questionable, for the following reason. Although 5 atmos. excess hydraulic pressure applied in the body of the liquid by a piston will eliminate cavitation under the given conditions, applying 5 atmos. of oxygen above the surface will force 5 times as much oxygen into solution (Henry's Law). Then when the instantaneous pressure is reduced 1 or 2 atmos. in the negative part of the sound wave cycle, oxygen will come out of solution and bubbles will be formed in the liquid. Therefore, it seemed desirable to investigate this question further, using a more positive means for eliminating cavitation: degassing the solution by boiling under vacuum for at least 15 minutes, prior to irradiation.

EXPERIMENTAL

A sample of high molecular-weight polystyrene,[6] obtained through the courtesy of Dr. S. G. Weissberg of

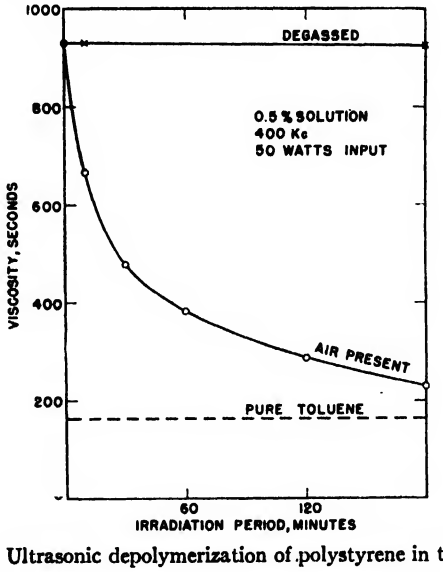


FIG. 1. Ultrasonic depolymerization of polystyrene in toluene.

[1] H. Mark, J. Acous. Soc. Am. 16, 183 (1945).
[2] K. Sollner, Chem. Rev. 34, 371 (1944).
[3] H. Freundlich and D. W. Gillings, Trans. Faraday Soc. 34, 649 (1938).
[4] By "cavitation" we mean the formation and violent collapse of small bubbles or voids in the liquid, as a result of pressure changes which occur, for example, in a sound wave.
[5] G. Schmid and O. Rommel, Zeits. f. physik. Chemie 185A, 97 (1939).
[6] Prepared by polymerization at room temperature for several months.

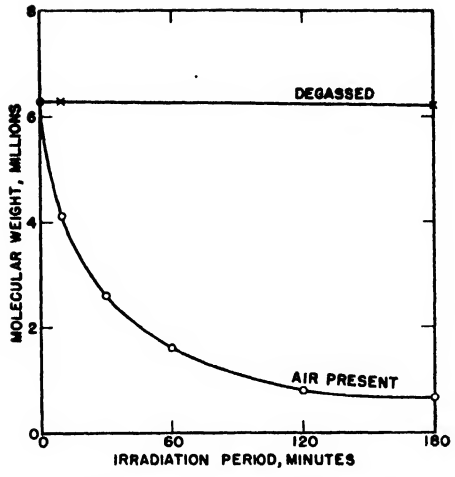


FIG. 2. Ultrasonic depolymerization of polystyrene in toluene.

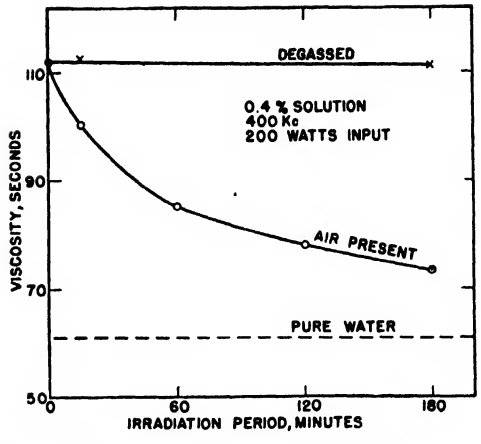


FIG. 3. Ultrasonic depolymerization of hydroxyethyl cellulose in water.

the National Bureau of Standards, was dissolved in toluene to give a 1 percent solution. A 50 ml sample of this solution, containing dissolved air, was irradiated[7] for 10 minutes (at a power level of 50 watts d.c. in the plate circuit of the output tube of the ultrasonorator) and then diluted to 100 ml. After this treatment the viscosity (measured at 30°C with an Ostwald-Fenske viscometer) was 665.7 seconds, compared to 929.3 seconds for an unirradiated diluted solution and 163.6 seconds for the pure solvent toluene. But when a similar sample was degassed first by boiling under vacuum, then cooled in the absence of air, irradiated, and diluted to 100 ml, no viscosity change was found (Fig. 1). Also, there were no cavitation bubbles in the degassed solution, although many were visible in the solution containing air.

Similarly, 3 hours irradiation gave only a very slight change (to 925.7 seconds) in a degassed solution, but a large viscosity decrease (to 270.9 seconds) in the solution containing air. Several intermediate points are also shown on Fig. 1. These viscosity decreases were permanent; no change could be detected after the solutions had stood for 20 hours.

The extent of depolymerization corresponding to the observed viscosity decrease was calculated from the equation

$$[\eta]=KM^a,$$

where M is the molecular weight, and $K=1.6\times10^{-4}$ and $a=0.66$ are two constants (obtained by extrapolation from data in the literature[8]) characteristic of toluene solutions of polystyrene which has been prepared by polymerization at 30°C, as was the present sample. The intrinsic viscosity $[\eta]$ is defined in terms of the viscosity of pure solvent η_0 and the viscosity η_c of a solution of concentration c:

$$\eta=\lim_{c\to0}\frac{\eta_c-\eta_0}{c\eta_0}=\lim_{c\to0}\frac{\eta_{sp}}{c}.$$

By means of successive dilutions, η_{sp} was evaluated for each irradiated solution at concentrations of 0.50, 0.20, and 0.10 gram/100 ml, and the results extrapolated to infinite dilution to give $[\eta]$. Figure 2 shows that the viscosity-average molecular weight decreases from about 6,000,000 to 600,000 in the ordinary solution, but in the degassed solution the decrease is negligible—of the same order of magnitude as the experimental error.

This result indicates that cavitation is responsible for the depolymerization. In order to determine whether the same is true in water solutions, analogous experiments were performed on a 0.80 percent aqueous solution of hydroxyethyl cellulose WS-1000 (obtained through the courtesy of the Carbide and Carbon Chemicals Corporation). Here again, ultrasonic irradiation

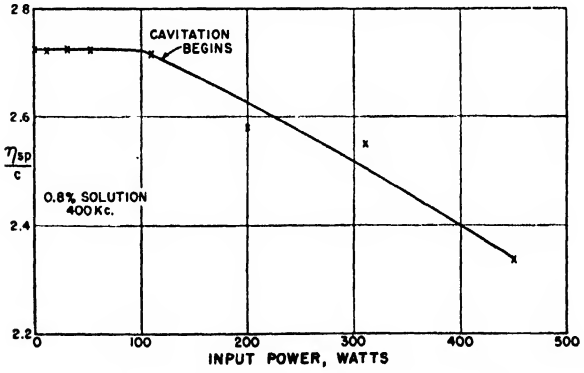


FIG. 4. Effect of input power on hydroxyethyl cellulose depolymerization.

[7] The container was a 50×400 mm test tube, closed at the top with a ground joint through which were sealed two gas-delivery tubes, one extending almost to the bottom of the container. This made it possible to change the nature of the dissolved gas in the solution. The ultrasonic waves used in these experiments had a frequency of 400 kilocycles.

[8] H. Mark in *Physical Methods of Organic Chemistry*, edited by A. Weissberger (Interscience Publishers, Inc., New York, 1945), p. 146; certain errors in this table were corrected for us by Dr. S. G. Weissberg.

for 3 hours at 200 watts input—followed by dilution to 0.4 percent—gave a very slight viscosity change (from 112.2 to 111.3 seconds) in degassed solutions, but a large decrease (to 73.6 seconds) in ordinary solutions; distilled water had a viscosity of 61.1 seconds in this viscometer. This is represented graphically in Fig. 3, together with several intermediate points.

An even more direct relation between cavitation and depolymerization is demonstrated in Fig. 4. It was previously observed that no cavitation bubbles could be detected in water irradiated at the lowest power levels, even with intense transverse illumination. However, bubbles could always be found at input powers above a given threshold, the value of which depended on the experimental conditions; in viscous solutions the threshold was higher and less sharp. In the present work, by subjecting successive portions of hydroxyethyl cellulose solutions to 5-minute irradiations at increasing power levels, it was found that no depolymerization occurred at the low levels where cavitation was absent. But just about at the power where cavitation began, depolymerization also began, and became greater as the power was increased.

It might be supposed that the observed degradation is due to oxidants which are known to be produced by irradiation of solutions containing nitrogen or oxygen.

TABLE I. Effect of dissolved gas on ultrasonic depolymerization.

Dissolved gas	Viscosity of 0.20% solution of polystyrene in toluene, after 30-minute irradiation	Viscosity of 0.40% solution of hydroxyethyl cellulose in water, after 15-minute irradiation
None	256.6 seconds	112.7 seconds
Air	177.6	100.3
Helium	175.0	100.0
Carbon dioxide	184.4	112.3

This view is invalidated, however, by the data in Table I; substantially the same amount of depolymerization occurs even when purified helium is the only gas present. Carbon dioxide also serves just as well in toluene, but not in water solution.

CONCLUSION

It is known that cavitation does not occur at moderate negative pressures in solutions which have been freed of dissolved gas. All concomitants of the ultrasonic wave, other than cavitation, are present in degassed solutions, but no depolymerization takes place. Therefore, it is cavitation which is responsible for the depolymerization caused by ultrasonic irradiation, in the polystyrene-toluene and hydroxyethyl cellulose-water systems.

A Three-Stage Electron Microscope with Stereographic Dark Field, and Electron Diffraction Capabilities

M. E. HAINE,* R. S. PAGE, AND R. G. GARFITT
Metropolitan Vickers Electrical Company, Ltd., Manchester, England
(Received August 2, 1949)

An electron microscope is described in which a range of magnifications between 1000 and 100,000 times can be covered continuously. This is achieved through the use of an intermediate projector lens.

The principles of design are briefly discussed and the construction of the microscope and the high frequency-operated 100-kv power unit is dealt with in some detail. The instrument is provided with objective apertures to enhance image contrast and to permit operation with dark field illumination.

Data is given relating to the electron optical system and it is shown how the instrument can be used to obtain electron diffraction patterns characteristic of small areas of a specimen under microscopic observation. The alignment procedure is briefly described and some typical micrographs are shown.

I. INTRODUCTION

FOLLOWING the work on the design and construction of a two-stage electron microscope Type E.M.2,[1] it was decided to undertake the development of a new instrument, the design of which would aim at obtaining the greatest possible simplicity and economy compatible with the best performance.

This paper describes the design principles evolved for the new instrument (Fig. 1) and explains the versatility in operation and also the mechanical, electrical, and electron optical details.

Previous electron microscopes have used a two-lens system, except that built by Marton[2] which used a three-lens system. The relative advantages of the two systems were considered and it soon became apparent that a three-lens system offered many advantages in providing a more compact design and a much greater versatility. Three-stage instruments have been de-

* Now with the Associated Electrical Industries Research Laboratory, Aldermaston, Berks, England.

[1] M. E. Haine, "The design and construction of a new electron microscope," J.I.E.E. 94, Part 1, No. 82 (October, 1947).

[2] L. Marton, "A 100-kv electron microscope," J. App. Phys. 16, 131 (1945).

veloped also coincidently with our own by le Poole[3] and Liebmann.[4]

A two-stage instrument consists of lenses separated by distances of three to four times the length of a lens. This means a long and hence not very rigid column. Ideally a column consisting of lenses, and space enough to allow for the necessary aligning mechanisms, would suffice. If we fix the ideal length per stage at a value l_0 and assume the lenses all have focal lengths f_0, then with n stages the magnification is given approximately by

$$M=(l_0/f_0)^n.$$

Thus,

$$n=\log M/\log(l_0/f_0).$$

For a maximum magnification of 100,000 times, $l_0=20$ cm, and $f=0.4$ cm, n becomes 3 to the nearest integer. The resulting three-stage column is 60 cm long compared with 250 cm for the corresponding two-stage instrument.

Further consideration shows that a much wider range of magnification for a given specimen setting is possible. For instance, whereas in a two-stage instrument the magnification can only be varied over a range of about five to one without changing the specimen position, in the three-stage instrument a range of nearly one hundred to one is possible. This further simplifies design as it obviates the necessity for an intermediate viewing screen for aligning purposes.

A further advantage results by virtue of the ease with which diffraction photographs can be obtained by imaging the diffraction pattern at the back focal plane of the objective lens on to the object plane of the second projector lens, using the first projector lens at unity magnification. The detailed mechanism involved is described in later paragraphs.

In addition to using a three-lens magnifying system, it was thought possible to reduce the over-all volume of the apparatus sufficiently to keep the "pump-down time" low enough to allow complete evacuation in two or three minutes and thus obviate air locks for specimen and photographic plate. This was considered particularly important in view of the difficulties already present in the mechanical stage design which are much aggravated if an air lock is necessary.

Our own past experience and that reported to us from other workers in this field made us decide to use an electron accelerating voltage of 100 kv. There seems little doubt that this represents an appreciable improvement on the 50 kv previously used, and yet it is evident that to increase the voltage to much over 100 kv involves considerable complication and expense.

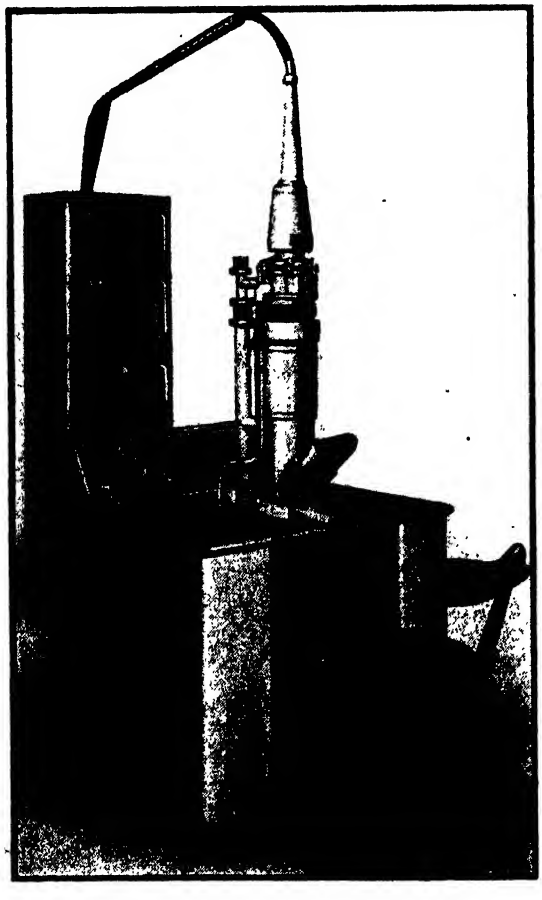


FIG. 1. Photograph of the E.M.3 electron microscope.

[3] J. B. le Poole, "A new electron microscope with continuously variable magnification," Philips Tech. Rev. 9, No. 2 (1947).
[4] G. Liebmann, "A new experimental electron microscope," J. Sci. Inst. 25, No. 2 (February, 1948).

II. DESCRIPTION OF MICROSCOPE COLUMN

The microscope column (Fig. 2) is built up in eight sections: The electron gun, condenser lens, specimen chamber, objective lens, intermediate alignment section, twin projector lenses, viewing chamber, and camera. Each unit is supported directly by the section underneath and vacuum sealed by rubber gaskets. Mechanical rigidity, a first consideration in electron microscope design, is maintained by the use of the type of rubber gasket seal used on the Type E.M.2 instrument which allows the opposing flanges to be clamped together in metal to metal contact.

A. The Electron Gun and Condenser Lens

The electron gun is totally enclosed, utilizing a screened cable entry bushing insulator (103) to conduct the high voltage lead (101) to the cathode. Thus no high voltage is exposed at the electron gun and the necessity for a cumbersome screening compartment is obviated.

The three-core high voltage cable is terminated in a three-contact plug assembly which fits into the central bore of the conical porcelain insulator and makes contact with the cathode assembly of the gun. A conical screen surrounds the whole insulator and serves to connect the cable screening direct to the main body of the microscope. The cathode assembly (111) is of more

or less conventional design, consisting of a tungsten hairpin filament surrounded by a biased cathode shield. The anode (117) has a central aperture to allow the beam to pass out of the gun and also a concentric ring of holes for pumping. The gun is surrounded by a lead-lined x-ray shield (118).

The electron gun is mounted directly on top of the condenser lens; no transverse alignment of the lens with respect to the gun axis has been found necessary.

The electron gun and condenser lens assembly is mounted on an aligning mechanism which allows it to be translated laterally or tilted about a point in the object plane for electron optical alignment. The mechanism employed is similar to that used in the Type E.M.2 instrument, but it has been reduced in size and made more rigid. The lateral movement is obtained by sliding two flat surfaces over one another with the adjusting screws (211) while the tilt is obtained by sliding two spherical surfaces centered at the specimen by means of adjusting the screws (212). Vacuum sealing is effected by a flexible metal bellows. The complete unit is mounted on top of the object chamber and sealed by a rubber gasket.

B. The Specimen Chamber and Mechanical Stage

The specimen chamber (Fig. 3) consists of a brass cylinder brazed to upper and lower sealing plates (302, 303). To provide access to the mechanical stage the front section of the cylinder is cut away to accommodate the door (304), which is sealed with a flat rubber gasket held to the door by an internal retaining plate. A swing catch pivoted at the center of the door engages in slots in the upper and lower flanges. Atmospheric pressure is sufficient to seal the door without addition of any auxiliary pressure, and the catch is provided merely to hold the door closed before evacuation commences. A pumping manifold connection is made at the back of the object chamber.

The object chamber is fixed directly on top of the objective lens by flange bolts, and sealed with a rubber cord gasket.

The object stage itself consists of a cup-shaped carriage (306) which slides directly on the upper face of the objective lens. These sliding surfaces are polished to reduce the friction so that a smooth movement is obtained, while two spring-loaded plungers hold the surfaces in close contact.

The specimen holder (309) is locked into the stereo block (310) by means of the finger-controlled screw (311), and is shown in detail in Fig. 4. The holder can be extracted from the stage through the chamber door with the aid of the special tweezers (Fig. 4) which engage with an internal groove in the holder.

The stereo block into which the specimen holder fits is mounted on two pointed pegs which rest in grooves in the base of the carriage, lying approximately in the plane of the object. The upper face of the block is cylindrically machined about an axis containing the

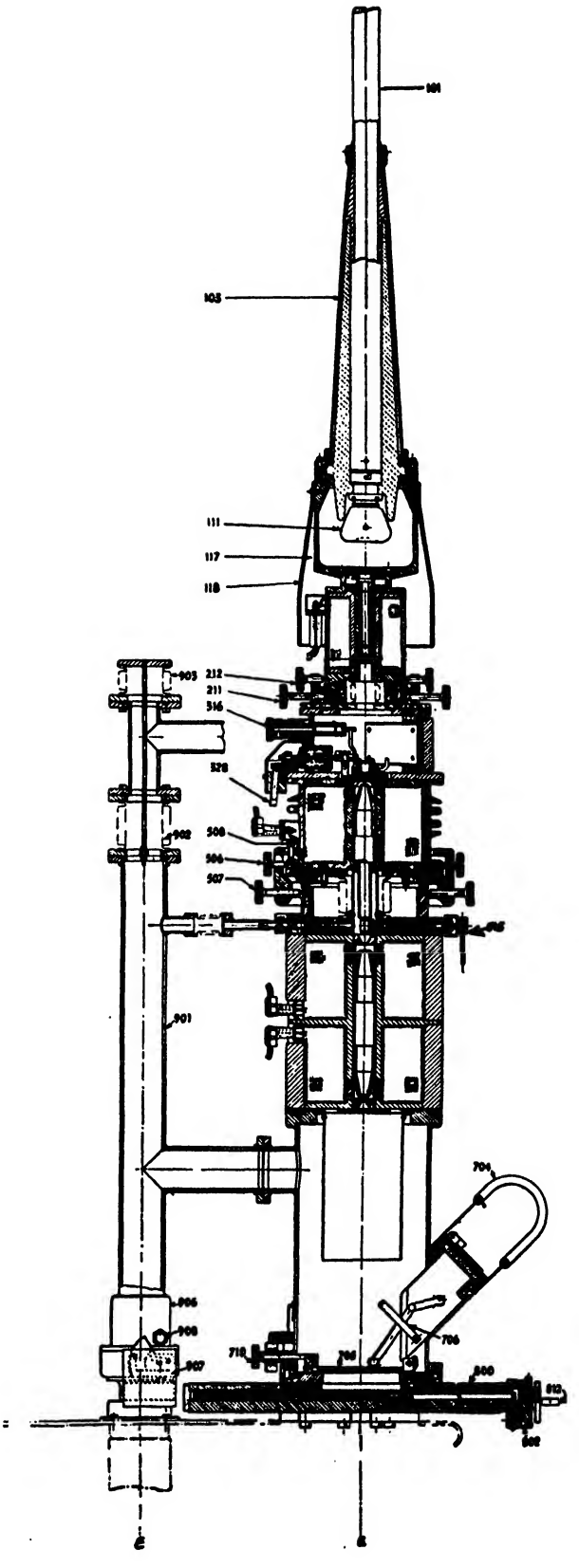


FIG. 2. Cross-sectional drawing of the microscope column.

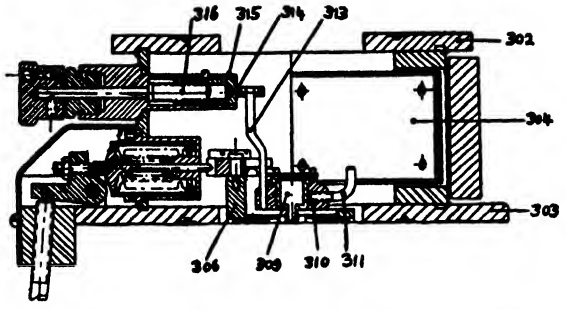


FIG. 3. Cross-sectional drawing of the specimen chamber.

points of the pegs. Leaf springs fastened to the carriage bear on this cylindrical surface, allowing the block to be tilted through a total angle of 10° about the axis in the object plane. Tilting is effected by a vertical lever (313) attached to the stereo block and engaging the tilting mechanism at the top of the specimen chamber. The tilting control consists of a lead screw (316), entering the vacuum through a rubber gland seal, which moves the block (314), keyed to the fixed tube (315), laterally. A rectangular opening in the block allows clearance around the lever so as not to interfere with the mechanical stage movement.

Motion in two directions at right angles is transmitted to the object stage through a lead screw and bell crank system from two knobs at hand level. Careful attention to the friction, controlling springs and rigidity of the operating mechanism gives smooth control to 0.1μ or better. The relatively small mass of the stage and close contact with the objective lens ensures thermal equilibrium between these being reached in a short time.

C. The Objective Lens and Intermediate Aligning Mechanism

The objective lens is of conventional design; the bore is made vacuum tight by a brazed copper spacer in the gap and is ground to give accurate cylindricity.

The pole pieces are asymmetric about the lens center, in order to bring the object focal plane to a position just above the top of the lens. A special grinding technique has been developed to ensure a minimum degree of eccentricity and ellipticity of the pole piece bores to minimize astigmatism. Further improvement in this direction is achieved by making the bore diameter as large as practicable with no increase in focal length, as by so doing the effect of residual mechanical and magnetic defects as well as the primary spherical aberration is reduced.

A focal length of 4 mm is obtained at 100 kv with the specimen plane slightly above the level of the upper lens face and a pole piece diameter of 2.5 mm. The pole pieces are finished by precision grinding to be an accurate push fit in the bore of the lens and can be removed for cleaning relatively easily.

The alignment section immediately below the objective lens provides two movements, one for translating (506) and one for tilting (507) the objective lens and illuminating system with respect to the projector lenses. The tilt takes place about a point in the object plane of the second projector lens. Each movement is controlled by a set of four self-locking screws.

D. Objective Apertures

The microscope has been fitted with two types of objective aperture; one being a relatively large aperture intended to give an increase in image contrast and to be easily inserted and removed. It takes the form of a hollow phosphor bronze cylinder closed at one end by a thin wall into which is drilled the aperture hole 100μ in diameter. The cylinder fits closely into the bore of the objective lens and can be inserted and removed by means of a special tool through the specimen chamber door. The operation takes slightly longer than that of changing the specimen. The plane of the aperture is located approximately at the back focus of the lens so that electron diffraction patterns cannot be obtained with the aperture in position.

The alternative type of aperture takes the form of an assembly which fits into the lower pole piece of the objective lens from the underside. The aperture, again located at the back focus of the objective lens, can be preset in lateral position by four grub screws and is centered before assembly of the objective lens by means of a low power microscope and "V"-block. This precentering allows the use of a smaller aperture. The assembly is thus semipermanent and precludes the use of electron diffraction or low magnification dark field imaging. Normal dark field images can be obtained, however, by tilting the illuminating system sufficiently for the direct beam to be intercepted by the aperture.

In addition, a special intermediate section has been developed for the instrument which carries an adjustable objective aperture assembly, details of which can be seen in the cross section of Fig. 5. The aperture is carried on the detachable end of a tube (551) which slides vertically inside a carrier tube (552) supported in a spherical seating (553) about 3 cm below the center of the objective lens. Thus, transverse movement of the lower end of the carrier tube will translate the aperture in the lens field. A main supporting tube (554) carries the spherical seating at the upper end and a cylindrical block at the lower end. Two fine-threaded

FIG. 4. The mechanical stage, specimen holder, and insertion tweezers.

screws through this cylinder allow adjustment of the position of the carrier tube. These screws extend to knobs outside the walls of the alignment section.

Vertical movement of the aperture tube is controlled by the fork (556) engaging with two pins in the lower end of the tube. The fork is fastened to a shaft projecting through the walls, and a knob on the end engages with stops suitably placed on the external wall.

The normal operating position for the aperture is about 4 mm below the lens center. When it is not required it can be raised to just below the level of the specimen plane so that the cone of electrons leaving the object area is unrestricted. When the specimen is removed, the aperture can be lifted even further so that it projects above the level of the lens face. This enables the cap containing the aperture disk itself to be removed or exchanged without having to dismantle the whole lens assembly.

The aperture used is a thin copper disk produced by Kodak. The hole sizes so far used are 30 and 60μ in diameter.

The lower plate of the aligning section carries a pumping port connected to the vacuum manifold, also another movable aperture system (515) inserted at the level of the first intermediate image. This is used in obtaining electron diffraction patterns as described in Section III. It consists of a cylindrical cap on the end of a shaft controlled by a knob. Two clearance holes in the cap allow the beam to pass unrestricted for normal operation while rotation of the knob through 60° or 120° brings two small apertures of 50μ or 250μ diameter, respectively, in line with the axis of the beam. A screw backstop is provided for initial alignment of these apertures.

E. The Double Projector Lens, Viewing Chamber, and Camera

The construction of the double projector lens is clearly shown in Fig. 2.

The yoke, copper spacers, and end plates are brazed together, giving a vacuum-tight bore. The lens winding is wound directly on the double bobbin so formed and the two halves of the outer iron shroud slide over the whole assembly. The bore is precision ground and sufficient accuracy is obtained to render individual alignment of the two lenses unnecessary.

The viewing chamber is provided with a single wide viewing port with a lead glass window and eyeshield (704). On the sides of the viewing port are fitted the controls for raising the fluorescent screen (705) and for positioning a magnifying lens (706) to aid correct focusing.

A pumping port connects the viewing chamber to the vacuum manifold and the top flange supports a cylindrical μ-metal screen. The viewing chamber is fastened on blocks mounted directly on the main cross members of the desk frame and the camera is strapped underneath

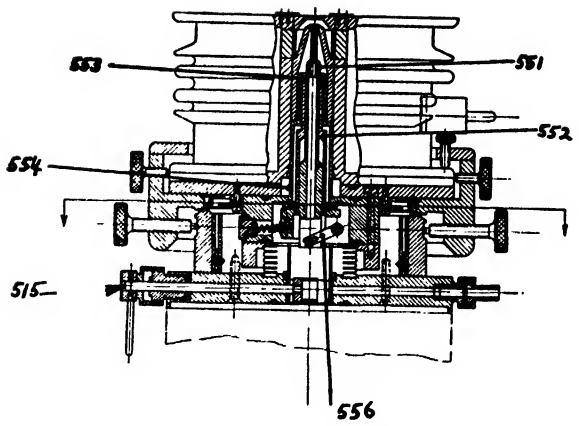


FIG. 5. Cross-sectional drawing of the intermediate alignment section with the adjustable objective aperture.

the lower flange, so that it can be removed without disturbing the rest of the microscope.

The camera consists of a shallow rectangular box with a transverse shaft operating the cassette tray through a pinion and rack. The door (802) is sealed by a flat rubber gasket and held in place by a spring clip. The drive for the racking pinion is brought out through a gland seal and bevel gears to knob (812) at the front right-hand side of the camera and a simple position indicating mechanism is incorporated.

The camera will take two $3\frac{1}{4}\times3\frac{1}{4}$-in. plates end to end, on which two, four, or eight micrographs can be taken. The frame size is selected by masks attached to the fluorescent screen; the appropriate mask is selected by the knob (710) at the rear of the viewing chamber.

F. Vacuum Equipment and Desk

The pumping equipment consists of a Metro-Vick Type D.R.I. two-stage rotary pump backing a Type 02 oil diffusion pump. The pumping system is manually controlled and provided with simple interlocks to prevent damage due to incorrect operation. The backing pump can be connected direct to the microscope column for rough pumping or to the diffusion pump for normal operation.

A simple fine side vacuum valve is incorporated between the diffusion pump and pumping manifold so that the microscope can be open to the atmosphere without shutting down the pumps. The over-all pumping time from atmospheric pressure is about two minutes.

The pumping manifold leading from the diffusion pump via the fine side valve has three connections to the microscope column (Fig. 2) as previously described; one at the viewing chamber, a second at the intermediate section, and a third at the object chamber. Flexible connections are made in the latter two to allow for the necessary alignment of the microscope column. The bending moment exerted on the column by air pressure on the top flexible connection (902) is com-

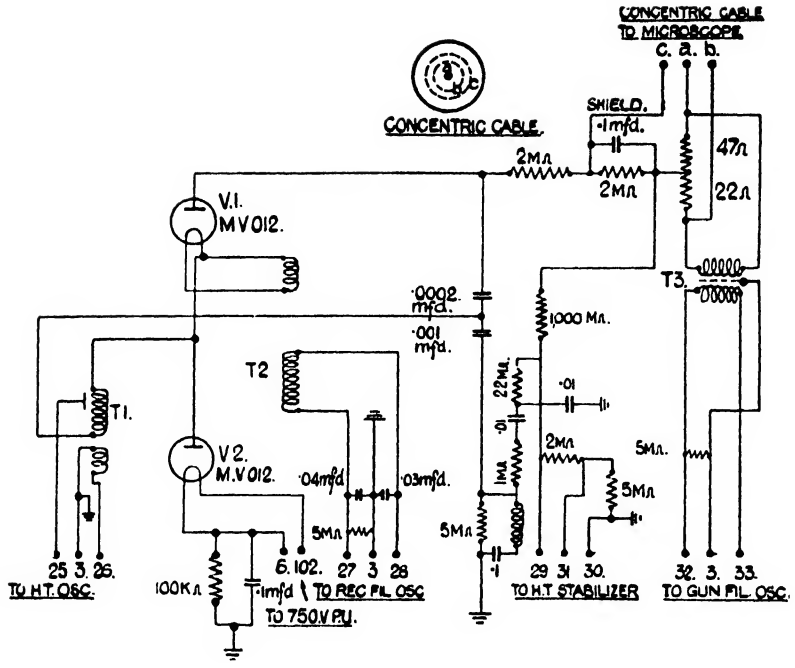


FIG. 6. Circuit diagram of the oil tank assembly.

pensated by a further flexible bellows (903) on the top of the manifold. The top plate sealing this bellows is mechanically connected to a spider in the lower end of the connecting bellows by a stay rod. Thus no resultant moment acts on the column.

The microscope desk (Fig. 1) is constructed from an angle iron framework with sheet metal paneling, and is mounted on rubber-bonded feet to insulate against vibration. A sloping panel at the rear of the desk displays beam current, lens current, and vacuum gauge meters, indicating lights and the a.c. mains circuit switches and fuses. The microscope controls are all grouped on a hinged panel just to the left of the column, while the lens current stabilizers are built on subpanels fastened to the control panel, providing easy access for maintenance.

All internal wiring takes the form of multiple cables or looms interconnecting each unit, terminated with multiple terminal blocks and spade end-coded leads.

G. The High Voltage Power Unit

This follows what may now be regarded as conventional practice for this type of instrument, the high voltage being obtained by rectification of the output of a resonant high frequency transformer fed from a valve oscillator.[5] Stabilization is obtained through negative feedback from a high resistance potential divider connected across the output circuit, via a d.c. amplifier, to the screen circuit of the oscillator valves, four 45-volt batteries providing the stable comparison voltage.

The complete unit contains three separate oscillators supplying power to the high voltage circuit, the high potential rectifier valve filament, and the electron gun filament. The operating frequencies are, respectively, 30, 50, and 50 kc/sec. All the oscillator and stabilizer components are grouped on a single rack-mounted chassis, while a second chassis contains the low voltage power supplies to the oscillators. The high voltage unit, which employs a Greinacher voltage doubler circuit, is housed in a brass oil-filled tank, and the electron gun supply is fed out through a shielded concentric cable. The voltage output of the set is varied by selection of the number of reference batteries in circuit, four 25-kv steps being provided. On the 25-, 50-, and 75-kv ranges, maximum output current is about 1 ma. and 0.4 ma. on the 100-kv range.

The electron gun filament power is remotely controlled from the microscope desk by variation of the anode supply voltage to the appropriate oscillator. Switching of the high voltage supply is accomplished by means of push buttons and relays, which also afford automatic overload protection. Figures 6 and 7 are circuit diagrams, respectively, of the high voltage and oscillator-stabilizer units.

III. ELECTRON OPTICAL PERFORMANCE

A. Illuminating System

The triode electron gun operates with an automatic bias derived from a variable resistor connected between the cathode and the cathode shield. For a given setting of the bias resistor there is thus a value of filament heating current which gives the correct lens action at the cathode shield aperture and produces an intensity maximum at the beam cross-over situated near the anode aperture.

[5] O. H. Schade, "Radiofrequency-operated high voltage supplies for cathode-ray tubes," Proc. I.R.E. 31, No. 4 (April, 1943).

Operating a triode gun under these conditions minimizes the effect of bad alignment of the filament in the shield aperture and also gives maximum brightness conditions. An image of the beam cross-over at approximately unit magnification can be formed in the object plane of the objective lens by the magnetic condenser lens, so that by operating the microscope at low power an image of the crossover can be projected on to the viewing screen. The cross-sectional diameter of the crossover determined in this way has a minimum value of about 0.005 cm, the actual value depending on bias setting and on the height of the filament in the shield. Under optimum conditions, the beam divergence from the crossover is approximately 0.03 radian. With a total beam current of 200 μa, the current passing through the condenser aperture, which subtends an angle at the cathode of 4×10^{-3} radian, is

$$200\times10^{-6}\times\left(\frac{4\times10^{-3}}{3\times10^{-2}}\right)^2=3.56\ \mu\text{a}.$$

The mean current density at the object is therefore

$$\frac{3.56\times10^{-6}}{\pi(2.5\times10^{-3})^2}=0.18\ \text{amp./cm}^2.$$

For adequate brightness on the final screen, a current density of 10^{-11} amp./cm^2 is required; hence the gun gives a sufficient intensity for operation at a magnification of $(0.18/10^{-11})^{\frac{1}{2}}=135{,}000$ times at critical illumination, or at an angular aperture of 5×10^{-4} radian, the visibly usable magnification is

$$1.35\times10^5\times\left(\frac{5\times10^{-4}}{4\times10^{-3}}\right)^2=2100\ \text{times}.$$

B. Lens System

Consideration of minimum aberration in the objective lens calls for the use of the shortest possible value of focal length. For a given pole piece design the minimum focal length is determined by magnetic saturation of the iron. Scaling up the dimensions of the pole pieces while maintaining constant field strength in the gap again shortens the focal length. Limitations here arise due to excessive increase in the excitation ampere turns required and also due to the fact that the lens field becomes so extensive that the back focus comes within the field, after which rays are deflected back toward the axis, and hence the focal length no longer decreases. A further advantage in scaling up the pole pieces results from the relative reduction in the astigmatism resulting from mechanical inaccuracies.

A further limitation in the focal length arises because of the difficulty of accommodating the object holder in

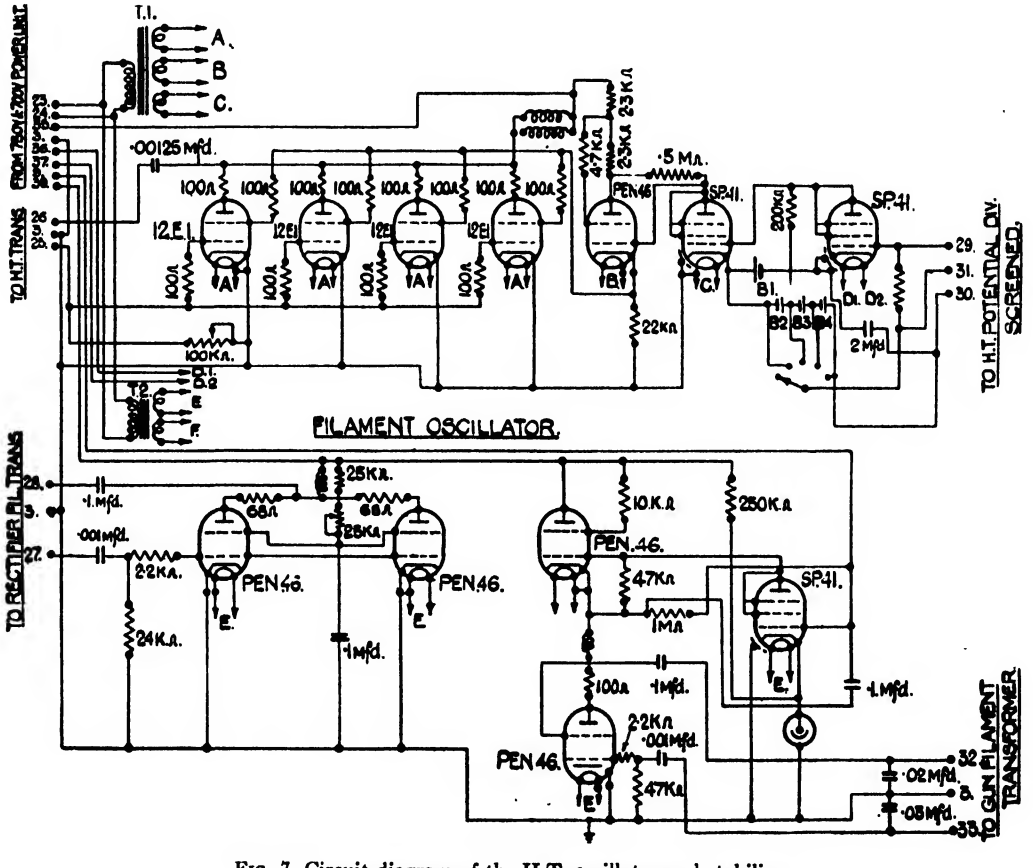


FIG. 7. Circuit diagram of the H.T. oscillator and stabilizer.

TABLE I. Magnification ranges.

ccelerating voltage	Low magnification undistorted	High magnification
100 kv	600–2000	2000– 60,000
75 kv	1000–3000	3000–100,000
50 kv	1500–5000	5000–100,000+

osition very close to the lens without severely tricting its movement. For this reason a focal length 4 mm is used, and the pole piece aperture chosen to as large as possible for saturation at the maximum itation. The maximum objective magnification is n about 35 times.

The three-lens system with the adjustable aperture vides a very flexible imaging system which can oper- under three distinct combinations of focal lengths. ese combinations provide for (1) high range magnifi- ion, (2) low range magnification, and (3) electron raction.

C. Magnification Range

The ranges of magnifications obtainable are shown Table I. On the high magnification range the two jector lenses are used in the normal way, that is, h producing inverted magnified images. The ray h for this condition is shown in Fig. 8(A). To avoid essive pincushion distortion† at the lower magnifica- ns, the second projector is operated at high power l gives most of the magnification (about 50× at) kv). For an 8-cm diameter image, the second pro- tor object field is then 0.16 cm in diameter, so that the bore of this lens can be made small without introducing distortion. Thus, for an over-all magnification of 2000 and first-stage gain of 35, the second lens must give unity magnification. The higher magnification can be obtained now by increasing the first projector power up to a maximum when it works at a focal length of about 4 mm, giving a stage magnification of 35× at 100 kv. This corresponds to an over-all magnification of 60,000×.

To obtain the low magnification condition, the first projector is used to form a reduced erect image by operating at low power. The ray path is shown in Fig. 8(B). The extent of the imaging beams at the projector lens aperture is now of the order of 0.3 cm so that the bore of the lens must be double that required by the high magnification condition in order to avoid barrel distortion. The wider bore is adopted though it slightly reduces the highest magnification obtainable.

D. Electron Diffraction

In this instrument an electron diffraction pattern corresponding to a small selected area of the specimen under examination can be formed immediately and without disturbing the specimen. This is achieved by using the first projector lens to form an image of the diffraction pattern existing in the back focal plane of the objective lens at approximately unit magnification in the object plane of the second projector lens. This image is then enlarged on to the final screen, or photographic plate. This method of obtaining diffraction patterns was originally used by Boersch.‡[6]

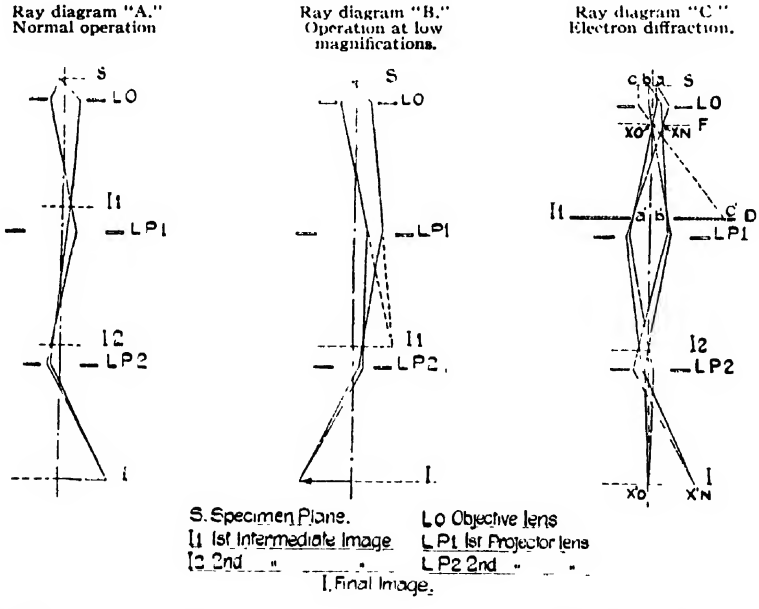


FIG. 8. Diagram of the ray paths.

† To insure negligible pincushion distortion, the area utilized in the object plane of the projector should be one-fourth of the e piece bore diameter.

‡ A similar result has been obtained by le Poole using a slightly modified arrangement incorporating an additional weak lens ference 3).

[6] H. Boersch, "On the primary and secondary image in the electron microscope II. Structure investigation by means of tron diffraction," Ann. d. Physik. 27, 75 (1936).

Under these conditions, the objective lens is used in the normal manner and therefore produces an intermediate magnified image of the object just in front of the first projector lens. A stop placed here will limit the field of the object observed in normal microscopic observation and will also, therefore, in the diffraction usage, limit the area of the object from which diffracted electrons are transmitted. The stop can be relatively large in diameter since the effective diameter is reduced by the magnification of the objective lens (35×). Thus, in practice, two stops 50 and 250μ in diameter are used, corresponding to fields of 1.5 and 7μ at the object. The ray path is shown at Fig. 8(C).

The maximum angle of diffraction allowed for is 0.1 radian. If the whole of the beam contributes to the intermediate image, the confusion disk diameter, given by $C_S\alpha^3M\cdot(C_S=0.5$ cm) becomes 175μ which is larger than the smaller stop. The effect of this when using the stop depends on whether the paraxial or marginal electron rays are focused at the aperture. Most of the intensity will normally be in the paraxial beam and if this is focused, then some of the energy from the outer rings will be lost and some energy from the outer rings of any crystals outside but near the defined area will appear. A better condition is for the intermediate image to be underfocused by such an amount as to allow all the outer ring electrons to pass through the aperture.

The diffraction pattern rings in the back focal plane of the objective will have a diameter given by

$$D=2\alpha_D f,$$

where α_D is the angle of diffraction and f the focal length of the objective lens. Toward the outside of the pattern some distortion is introduced by spherical aberration. This has the effect of reducing the diameter of the outer rings by an amount $C_S\alpha_D^3$. For a maximum angle of diffraction of 0.1 radian and a focal length of 0.4 cm, the ring diameter becomes $D=2\times0.1\times0.4$ cm $=800\mu$ and the distortion is $\Delta D=0.5\times(0.1)^3$ cm $=5\mu$ or a little under one percent. A simple correction can be applied for this error.

The final projector lens provides sufficient magnification to enlarge the diffraction pattern to a convenient size which is just too small for distortion in this lens to be significant.

The diffraction rings are made up of overlapping images of the electron source and therefore to give finely resolved rings the source must be small. If we operate the instrument under conditions of critical illumination we should expect to get no diffraction at all, as incoherent radiation is irradiating the object. In fact, a fair degree of coherence results from the imperfections in the cathode and condenser lenses, and even if this were not so, the Airy disk produced by the condenser lens aperture would give sufficient coherence for some pattern to result; the resolution would, however, be very poor.

If the condenser lens is switched off the electron gun virtual cathode is the source and this is demagnified about 50× by the objective into the back focal plane giving a ring width of about 1μ. To improve upon this, the condenser lens can produce an image of the source demagnified about 10×. The image is closer to the objective and hence the objective demagnification is reduced to 20× giving an over-all demagnification of 200× or an image size of about 0.25μ. This for the outer rings corresponds to a resolution of 1 in 3000 which is adequate.

Spherical aberration of the objective lens will have a further detrimental effect on the outer rings of the diffraction pattern. The diameter of the disk of confusion in the back focal plane is given by

$$\delta_S=d(C_S\alpha_D^3)=3C_S\alpha_D^2d\alpha.$$

Here α_D is the angle of diffraction and $d\alpha$ is the angle subtended by the intermediate diffraction aperture at the diffraction pattern. For the 250μ stop, $d\alpha$ becomes 0.002 radian, and if $\alpha_D=0.1$ radian, then $\delta_S=3\times0.5\times10^{-2}\times2\times10^{-3}=0.3\mu$.

If this error is added directly to that due to the finite source size, the resolving power of the system becomes 1 in 1500 with an improvement toward the center of the pattern.

E. Electron Optical Alignment

The alignment procedure employed with this instrument is very similar to that used on the Type E.M.2 microscopes.[1,7] The system is first aligned as a two-stage

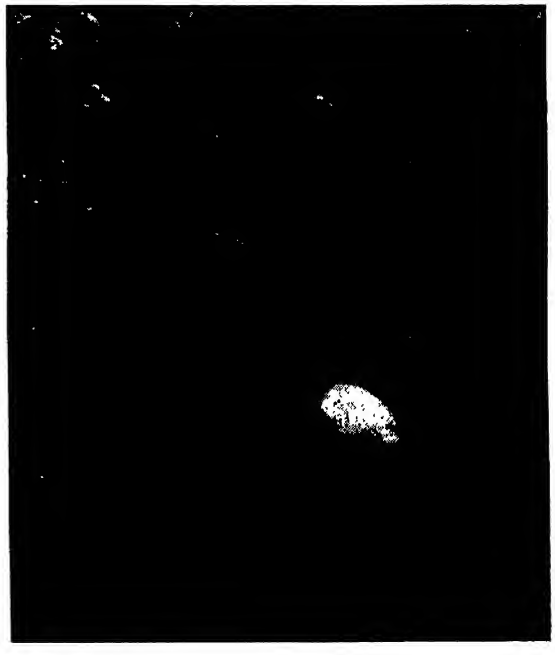


FIG. 9. Staphylococcal bacteriophage and staphylococcus, shadow cast at 4:1. (32,000 times, 75 kv.)

[7] M. E. Haine, "The electron optical system of the electron microscope," J. Sci. Inst. 24, No. 3 (March, 1947).

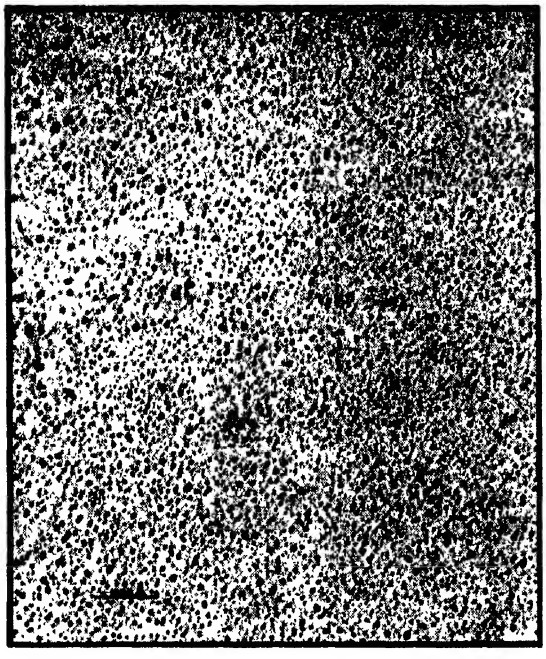


FIG. 10. Evaporated palladium gold. (72,000 times 75 kv.)

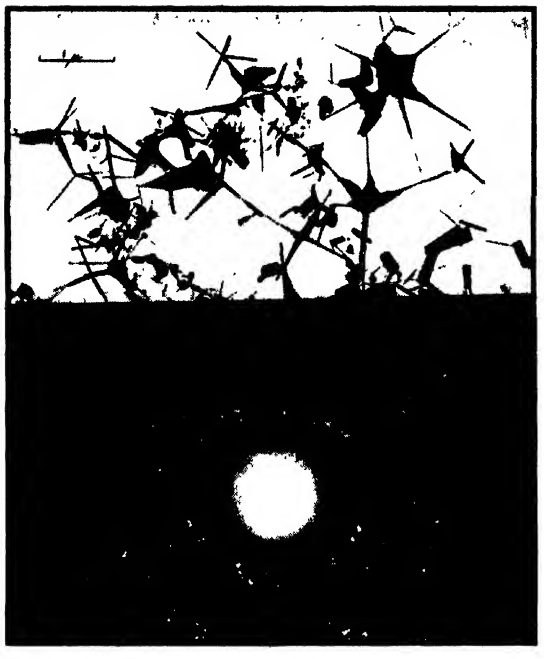


FIG. 11. Electron diffraction pattern of zinc oxide crystals, 100 kv.

microscope with the first projector lens switched off, using reversal of the objective lens current to locate the lens axis.

The first projector lens is then switched on at low power to give an image of the back focus of the objective lens in the image plane of the final projector (electron diffraction condition). Under these conditions the spot on the final screen will be displaced from the center if the first projector lens is not aligned with the axis of the objective lens. This is corrected by use of the intermediate section tilt. As this tilt is arranged to occur about the center of the second projector lens, the alignment of this lens and the objective is not upset. However, if the tilt required is large it may be necessary to repeat the initial adjustment.

The magnification can then be raised to 10 or 20,000 times and final adjustment of the gun tilt carried out.

CONCLUSION

In operation, we have found this instrument to give the promised flexibility and simplicity. The absence of air locks proves no disadvantage; the specimen or photographic plates can be changed in about three minutes while such operations as changing an aperture or a filament take very little longer.

In regard to resolving power, the average value obtained during a series of tests with heavy metal particles of about 25A diameter was approximately 35A with a best value of 25A. Edge contours have been recorded with half-widths of about 15A while spacings of as low as 12A between maxima and minima in Fresnel fringe systems have been obtained. These results indicate that a resolving power of better than 50A should be obtained under normal operating conditions.

Figures 9 and 10 show typical micrographs while Fig. 11 shows the micrograph and diffraction pattern, obtained from a field 7.0μ in diameter, of zinc oxide crystals.

ACKNOWLEDGMENTS

The authors wish to express their appreciation of the help afforded them in the development of this microscope by Mr. R. W. P. Tostevin, Mr. H. M. V. Young, and Mr. H. Nuttall, also to Mr. G. A. Barnett for his collaboration on the mechanical design.

In conclusion, the authors express thanks to Mr. F. R. Perry for his encouragement in this development, to Sir Arthur P. M. Fleming, Director, and to Mr. B. G. Churcher, Manager of Research Department, Metropolitan-Vickers Electrical Company, Ltd., for permission to publish the paper.

Letters to the Editor

Comments on the Use of Latex Spheres as Size Standards in Electron Microscopy*

CHARLES H. GEROULD
Magnesium Laboratories, The Dow Chemical Company, Midland, Michigan
October 19, 1949

BACKUS and Williams[1,2] have reported the usefulness of Dow latex 580G, lot 3584, as a standard in the determination of magnification and shadow-casting angle in electron microscopy. They attributed to these latex spheres a size of 2590±25A. Since the unique value of this material became apparent, about 225 small samples have been distributed to laboratories in this country and abroad. Seventeen of these laboratories have made determinations of the average particle size and have submitted their results which have been tabulated in Table I. It is noted that these determinations include six different techniques; electron microscopy alone, combination of light and electron microscopy, light microscopy alone, interferometry with electron microscopy, light scattering and ultracentrifugation. It appears significant that thirteen of the seventeen determinations show an average size of 2588A with a range from 2520 to 2630A. The four other determinations range upward to 3200A.

It should be mentioned that more recent light scattering studies by the two laboratories listed have indicated that the procedure used is not reliable.

It has been reported[3] that the size of this latex may be affected by variations in the procedure of specimen preparation, such as degree of dilution, size of drop added to specimen, drying speed, etc. In order to check this possibility, sixteen experiments were performed in which the dilution of the 40 percent latex with water was varied between 1:250 and 1:4000, the size of the drop added to the specimen and the drying times varied to reasonable extremes, dilutions made from a drop of latex which had been allowed to nearly dry before diluting, and dilution of the settled and caked material found in the bottom of the latex bottle. The results indicated that none or, at the most, very few significant changes took place. Twelve experiments showed variation of less than 1 percent from the mean, two between 1 and 1.5 percent and two between 2 and 3 percent. These latter two experiments may be significant or they could be the result of instrumental fluctuation. However, in both of these cases, the 2 to 3 percent variation was a decrease in size as a result of the partial drying of a drop of latex before dilution, a procedure rarely encountered in actual practice.

TABLE I. Determinations of particle size of latex 580G, lot 3584.

Contributor	Affiliation	Average size in angstroms	Estimated accuracy in angstroms	Experimental procedure: Instruments used	Reference specimen	Type of replica	Remarks
R. C. Backus R. C. Williams	Univ. of Mich.	2590	± 25	LM-EM[b]	Latex on glass fibers	None	
G. H. Bergold	Can. Dept. of Agri.	3100	±100[a]	LM-EM	Glass fibers	None	Instrumental[c] Calibration
J. S. Boadway	Shawinigan Chem., Ltd.	2760	± 60	LM-EM	15,000 lines/in. Grating+Latex	Collodion	Replica checked using LM
W. B. Dandliker	Harvard Med. Sch.	2590		Light scattering	—	- -	
J. L. Farrant A. J. Hodge	Australian Council for Sci. and Ind. Res.	2560	± 25	EM	Glass fibers	None	Diameter of fibers determined by interferometric methods
E. F. Fullam	General Electric Co.	2590	± 50	LM-EM	30,000 lines/in. Grating+Latex	Parlodion-SiO_2	Replica checked using LM
C. H. Gerould	Dow Chemical Co.	2630	± 50	EM	15,000 lines/in. Grating+Latex	Polystyrene-SiO	
G. H. Hass	Engrs. Res. and Dev. Labs.	2620	± 40	EM	30,000 lines/in. Grating+Latex	Silver-SiO	
A. C. Jenkins	Linde Air Products Co.	2590	±130	EM	30,000 lines/in. Grating+Latex	Collodion	
H. Kahler	Nat. Inst. of Health	2600	± 65	EM	20,000 lines/in. Grating	Unknown	
R. A. Kern S. F. Kern	Eli Lilly and Co.	2580	± 50	LM-EM	Diatoms	None	Instrumental[c] Calibration
M. Martin	Godfrey L. Cabot, Inc.	2610	±130	EM	15,000 lines/in.	SiO	Instrumental[c] Calibration
O. W. Richards	American Optical Co.	3210	±240	Phase Contrast Light Microscope	—	—	Bright contrast using white light. Wet sample
E. H. Rowe	B. F. Goodrich Co.	2610	± 30	LM-EM	Opening in screen	None	Instrumental[c] Calibration
G. D. Scott	Univ. of Toronto	2500–3000	± 50[a]	LM-EM	15,000 lines/in. Grating+Latex	Formvar	Replica checked using LM
D. G. Sharp	Duke Univ.	2520	± 25	Ultracentrifuge	—	—	Particle Density—1.054±0.001
R. S. Spencer	Dow Chemical Co.	2560		Light scattering	—	—	

[a] On any one mount.
[b] LM—Light Microscope; EM—Electron Microscope.
[c] Instrumental Calibration—EM calibrated for magnification with specimen mentioned. Latex size determined using this calibration.

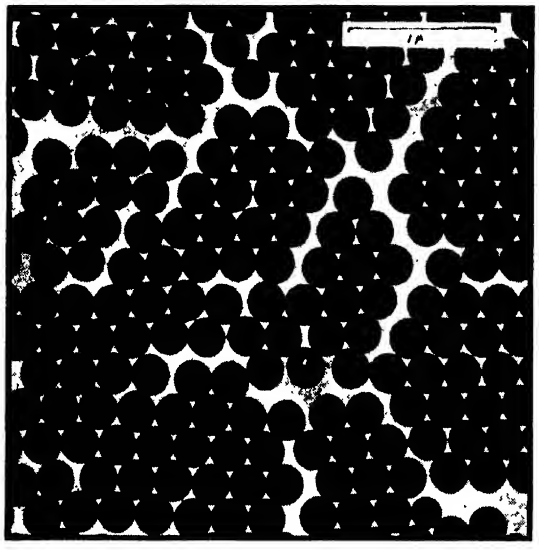


FIG. 1. Electron micrograph of single layer, latex particle array.

Since this unique batch of latex is of extremely limited quantity, the precautions in handling the material should be emphasized. Agglomeration is a danger and may be prevented by avoiding freezing, chemical contamination or change in pH. Agglomeration as a result of the evaporation of water from the undiluted latex is a hazard which is acute, especially in the 5 cc vials in which much of the latex has been distributed. Several laboratories have reported that their latex has completely dried even though the caps were apparently tight. It is suggested that either the bottles be periodically checked and kept filled to the top by additions of distilled water, or the whole sample be diluted 5:1 or 10:1.

The danger of bacterial contamination is not generally recognized. Experience has shown that certain bacteria can thrive and multiply in the latex resulting in eventual complete ruin of the material. This is difficult to combat, but special handling precautions and possibly refrigeration above 0°C will help.

As mentioned previously, a study was made on the effect of dilution ratio on particle size and no variation was found. However at dilutions of 1:500 and less, considerable fusing of the spheres resulted. In most cases, this fusing was slight near the wires of the supporting screen and severe in the center of the openings. An explanation may be that the surface active agents are present to such a concentration in the 1:500 or less dilutions that during drying, these agents are further concentrated to a point where they act as a plasticizer and cause fusing. It should be pointed out that the use of an atomizer may appreciably decrease this danger.

The danger of overheating in the electron microscope appears to be of importance only at the high intensities found with the biased gun. Fusing and distortion can result under such conditions. A much more real danger is found as a result of the specimen contamination which occurs under electron bombardment. These latex spheres make an ideal specimen for studying this phenomenon. Particles commonly show an increase in diameter of 4 percent to 5 percent after 20 minutes bombardment with a 300 microampere beam from a non-biased gun.

R. A. Kern and S. F. Kern† have recently discovered increases in the diameter of latex spheres as a result of shadowing, which cannot be explained by the mere thickness of the metal film on the particles. They have tentatively explained the size variations as being the result of an electrostatic charge on the unshadowed spheres which is removed by metal shadowing, providing it grounds the spheres. The possibility of such an effect should be carefully considered when these spheres are used in applications where shadowing is necessary.

Two electron micrographs are included which illustrate the particle arrays which are possible with spheres having the uniformity of diameter found in this material. Figure 1 shows a single layer particle array and Fig. 2 shows a multiple layer array in

FIG. 2. Electron micrograph of multiple layer, latex particle array. Shows areas of 1, 2, 3 and 4 layers of spheres.

which 1, 2, 3 and 4 layers of spheres can be seen. It is obvious that such arrays would be impossible without a high degree of uniformity of the spheres involved. One sphere of slightly different size will completely destroy the symmetry of the area around it.

At the present time, there are still some samples of this particular batch of latex 580G available to electron microscopists. They may be obtained by contacting the author.

* Condensed from paper given at the October, 1949, meeting of the Electron Microscope Society in Washington, D. C.

[1] R. C. Backus and R. C. Williams, J. App. Phys. 19, 1186 (1948).

[2] R. C. Backus and R. C. Williams, J. App. Phys. 20, 224 (1949).

[3] G. D. Scott, J. App. Phys. 20, 417 (1949).

† To be submitted to J. App. Phys. for publication.

Erratum: Relationship between Gough-Joule Coefficients and Moduli of Vulcanized Rubbers

[J. App. Phys. 20, 526 (1949)]

F. S. CONANT, G. L. HALL, AND G. R. THURMAN
Chemical and Physical Research Laboratories, The Firestone Tire and Rubber Company, Akron Ohio

IT has been pointed out by Dr. Sidney M. Katz of Textile Research Institute Laboratories that the thermodynamic reasoning leading to Eq. (8) of the above article is faulty, although the equation itself is valid. A correct thermodynamic derivation is possible, as Dr. Katz has demonstrated in a private communication. In view of the relatively simple mathematical derivation of Eq. (8), however, any involved thermodynamic approach appears unnecessary.

Frictional Loss in Visco-Elastic Substances

B. Gross
National Institute of Technology, Rio de Janeiro, Brazil
November 22, 1949

In two previous papers we have developed the phenomenological theory of linear visco-elastic behavior.[1] In this theory the elastic loss was associated exclusively with the relaxation mechanism which is also responsible for the frequency dependence of the dynamic modulus. On this basis the loss is uniquely determined when the dynamic modulus function is known. It seems, however, to be likely that another cause for elastic loss also exists, i.e. a "true" frictional mechanism which leads to a loss varying independently of the visco-elastic loss. The corresponding frictional force is assumed to be proportional to the rate of change of deformation and the proportionality factor, which is called frictional coefficient, is considered to be constant.[2] It might be instructive to examine experimental data for the presence of such a constant friction. The new loss component will be of importance at very high rates of deformation (or high frequencies). Therefore, the mathematical treatment now will have to take into account inertia forces too; this means that it must be based on the wave equation. But it will be sufficient to consider the unidimensional case only, because in the recent methods for the determination of the visco-elastic constants by means of sound waves the samples have the form of long thin strips which for all practical purposes can be considered as unidimensional ones.[3]

For a linear strip of material the unidimensional wave equation can be written in the form

$$q\rho\frac{\partial^2 u}{\partial t^2}dx=k, \tag{1}$$

where u is the displacement from equilibrium position of a cross section, ρ the density, and k the resultant force acting on a volume element of the sample, i.e. the difference of forces between x and $x+dx$.[4] The force now contains 3 components: (a) A component which corresponds to Hooke's law and contains the instantaneous elastic modulus E. (b) A component which takes account of the relaxation mechanism and contains the relaxation function $\bar{\varphi}(t)$. (c) A component which represents the frictional force and contains the constant frictional coefficient γ. Thus k is written

$$k=qE\frac{\partial^2 u}{\partial x^2}dx-q\int_{-\infty}^{t}\frac{\partial^2 u(x,\tau)}{\partial x^2}\bar{\varphi}(t-\tau)d\tau dx+\gamma q\frac{\partial^3 u}{\partial t\partial x^2}dx. \tag{2}$$

(1) and (2) together constitute an integro-differential equation which replaces the simpler equations of the former papers. While the general solution is involved, a particular solution of the form

$$u=u_0\exp(i\omega(t-x/c)-\alpha x) \tag{3}$$

is easily seen to exist; it represents a linear wave traveling from $x=0$ to $x=\infty$, of frequency ω, phase velocity c, and attenuation factor α. Substitution of (3) into (1) and (2) gives

$$(E-\bar{A})+i(\omega\gamma+\bar{B})=\frac{\rho\omega^2}{(\omega/c+i\alpha)^2}, \tag{4}$$

where $\bar{A}$ and $\bar{B}$ are, respectively, the storage factor and the loss factor already given in II:

$$\bar{A}(\omega)=\int_0^\infty \cos\omega\tau\bar{\varphi}(\tau)d\tau, \tag{5a}$$

$$\bar{B}(\omega)=\int_0^\infty \sin\omega\tau\bar{\varphi}(\tau)d\tau. \tag{5b}$$

Equation (4) is formally identical with Nolle's Eqs. (5). Introducing the complex elastic modulus $E^*=E_1+iE_2$ there is

$$E_1=E-\bar{A} \tag{6a}$$

$$E_2=\gamma\omega+\bar{B}. \tag{6b}$$

According to Eqs. (6), the Hookean mechanism contributes only to the real part of E^*, the frictional force to the coefficient of the imaginary part, and the relaxation mechanism to both of them. Since $\bar{A}(\infty)=0$, the instantaneous elastic modulus is obtained from a plot of E_1 *versus* frequency as the high frequency limit of E_1. $\bar{B}$ is 0 for both $\omega=0$ and $\omega=\infty$; $\gamma\omega$ is linearly increasing with frequency. Therefore, at very high frequencies where E_1 has become practically constant, E_2 should increase linearly with frequency. From the slope of the loss factor curve at high frequencies one then may read γ. The relations between $\bar{A}$ and $\bar{B}$ were already discussed in II. It will be interesting to calculate $\bar{B}$ from data of $\bar{A}$ and to see whether the result still leaves room for a component $\gamma\omega$.

[1] B. Gross, J. App. Phys. **18**, 212 (1947); **19**, 257 (1948), in the present paper quoted as I and II.
[2] A term of this type *without* any assumption about the frequency dependence has already been introduced by A. W. Nolle, J. Acous. Soc. Am **19**, 194 (1947); J. App. Phys. **19**, 753 (1948).
[3] See R. S. Witte, B. A. Mrowca, and E. Guth, J. App. Phys. **20**, 481 (1949). D. G. Ivey, B. A. Mrowca, and E. Guth, J. App. Phys. **20**, 486 (1949).
[4] For this deduction see G. Joos, *Theoretical Physics* (G. E. Stechert, New York, 1934), p. 176.
[5] A. W. Nolle, reference 2; see A. Gemant, Trans. Faraday Soc. **31**, 1582 (1935); Physics **6**, 363 (1935).

Addendum: Interpretation of X-Ray Patterns of Cold-Worked Metal

[J. App. Phys. **20**, 885 (1949)]

B. L. Averbach and B. E. Warren
Massachusetts Institute of Technology, Cambridge, Massachusetts

In a letter under this title the next to the last sentence applied the Fourier coefficient method to the particle-size problem. "For a substance in which there is particle-size broadening only, the initial slope of the A_m *vs.* m curve gives an average particle size and the second derivative of the curve gives directly a size distribution." Unfortunately at the time of writing we were not aware that this same result had already been obtained by E. F. Bertaut[1] in France.

[1] E. F. Bertaut, Comptes Rendus **228**, 492 (February, 1949).

Bibliography on Offshore Petroleum Developments

E. N. Kemler
Southwest Research Institute, San Antonio, Texas
November 17, 1949

The Division of Oceanography and Meteorology of Southwest Research Institute recently published the first Bibliography on Offshore Petroleum Developments. Intended as a practical and useful guide, the Bibliography is a public service presentation of the Institute available to any organization, company, or individual whom it might benefit. Requests for copies should be directed to The Division of Oceanography and Meteorology, Southwest Research Institute, 312 Oil and Gas Building, Houston, Texas.

New Books

Modern Arms and Free Men

By Vannevar Bush. Pp. 273, Simon and Schuster, New York, 1949. Price $3.50.

Dr. Vannevar Bush has probably had a greater influence on *applied physics* than any other scientist of our time. His invention of the differential analyzer and its use for the solution of non-linear equations initiated the great developments in electrical calculating machines. He is famous for his leadership of the Carnegie Institution of Washington, where he has had a direct hand in the success of the 200-in. Mount Palomer telescope. He was chairman of the National Advisory Committee for Aeronautics during the critical pre-war years. But his greatest contribution is his wartime leadership of a substantial fraction of all U. S. scientists, as Director of the Office of Scientific Research and Development. In this position Dr. Bush's leadership made itself felt not only in the highest councils in Washington, but more importantly, through all levels of the vast OSRD organization, so that even the youngest men knew instinctively that all worth while developments would find effective use.

Dr. Bush gives us a very clear account of the continuing menace of the submarine, the combating of which remains the most pressing problem facing the navy of any maritime power in its mission to keep the sea lanes open for transport. In the first World War and again in the second, the German submarine held the outcome in doubt and it cost us heavily to overpower it. The technical development and tactics of sonar, radar, the sono-radio buoy, and the magnetic airborne detector, as well as greatly improved ship and aircraft-launched ordnance, finally turned the tide against the U-boat. But pro-submarine developments that came too late to re-establish the submarine's supremacy—the schnörkel, the long-range homing torpedo, and new types of fast, deep-traveling submarines are now in the hands of potential sea commerce raiders. The added fact that the submarine, along with the merchant ship, may well be the agent for delivery of the atomic bomb accounts for Dr. Bush's current insistence on vigorous anti-submarine developments.

The atomic bomb naturally occupies a central position in this book. In his considered judgment, Dr. Bush evaluates this as a weapon of extreme power but not absolute in the sense that it makes other modes of warfare obsolete. Gone, in his analysis, is the abject fear which paralyzes clear thinking and wise planning, and gone also are the overestimates which are equally fatal.

The atomic bomb presents a grave problem in its continuing strain on the peacetime economy of the nation. This is partly compensated for by the fact that of all war materials it is probably the best suited for stockpiling purposes and in addition holds potentially great possibilities for peacetime applications. In contrast to the submarine menace, the defense against the atomic bomb seems to Dr. Bush to have great promise in the future. The development of guided missiles with proximity-fused ordnance as well as the effective use of radar warning and attack systems backed up by jet fighters may well spell the doom not only of long-range atomic bombing, but also of all forms of mass bombing against an alert and fully equipped nation. That these defensive methods may be deciding factors is suggested by the experience in the Schweinfürt raid, and also in the late but very effective introduction of the proximity-fused shell for the defense of London and the battle of Antwerp. Guided missiles will thus not only extend the offensive range of artillery and the fire power of aircraft, but may even be decisive in dooming the future effectiveness of large naval ships, and in defending great cities. The use of guided missiles on intercontinental raids Dr. Bush rules out on the basis of their excessive cost as against other forms of attack.

Proper planning for a national emergency Dr. Bush emphasizes as the theme of this book. This includes a clear recognition that cost, in its broadest sense of the effective use of the manpower and industrial and scientific capacity of a nation, is the crucial element in war. He gives brief but clear examples of how the magnetron tube and early forms of radar and sonar were available to the British and ourselves but not to the Germans at the outbreak of the war. Furthermore, the fact that wartime developments in Germany failed not only to produce an atomic bomb, but more significantly that they failed to provide jet fighters in quantity for defense against bombing shows the fatal consequences of a system which ignores the power of free scientific inquiry. It is hoped that peacetime planning, and especially education, will fully realize the emphasis placed here by Bush, and elsewhere by Conant, for the necessity of proper educational opportunities for all gifted youth in the nation, regardless of their financial status. It may well be fatal to the national safety to waste their talents.

This is one of the most important books since the war. It should be required study for every person holding even a minor position in national planning, and it is to be hoped that it may be brought to the attention of many young people, some of whom may someday find themselves over night in positions of great responsibility in a national emergency.

R. S. Shankland
Case Institute of Technology

Journal
of
Applied Physics

Volume 21, Number 3 March, 1950

British Journal of Applied Physics

IT is a pleasure to welcome a new journal to the fold of physical literature. In January of this year there appeared Volume One, Number One, of the *British Journal of Applied Physics*. This new journal is published by the Institute of Physics in Great Britain under the editorship of Dr. H. R. Lang, the Secretary of the Institute. In a foreword in this first issue, Dr. T. C. Toy, President of the Institute of Physics, mentions the interest which the Institute has had in applied physics since 1918. The first fruit of this interest was the inauguration in 1923 of a series of monographs and lectures under the general title of "Physics in Industry." This was followed by the establishment of the Institute's *Journal of Scientific Instruments*. Now the parts of this journal related particularly to the new applications of physics and new developments have been separated and published together as the *British Journal of Applied Physics*.

The first issue of this new journal sets a high standard indeed. The two leading articles are of a general nature, one entitled "A Scientific Education" by Sir Philip Morris, and the other on "Some Chemical and Physical Properties of Rubber" by J. Moore. There then follow five original contributions on the Development and Deionization Time of Heavy-Current A.C. Arcs, Measurement of Opacity and Reflectivity for Printing Papers, Measurement of the Coefficient of Internal Friction of Solid Rods by a Resonance Method, The Application of Multiple Factor Analysis to Industrial Test Data, and a Note on the Interpretation of Multiple Factor Analysis. In addition, this issue reviews several new books of particular interest to applied physicists.

The new journal is printed on a high quality coated stock which allows excellent reproduction of halftones. The page size and general appearance are similar to that of the *Journal of Scientific Instruments*. A small number of advertisements are also included.

There is an interesting and significant parallel in the development of applied physics in Great Britain and in the United States. As mentioned, the Institute of Physics was established in Great Britain in 1918. In 1931 the several national organizations devoted to physics and its applications in this country established the American Institute of Physics. The first publication to come under the direct sponsorship of the American Institute was the *Review of Scientific Instruments*. An Advisory Council on Applied Physics was established, and the theme of the large conference celebrating the fifth anniversary of the founding of the American Institute was "Physics in Industry." Shortly thereafter, in 1937, the *Journal of Applied Physics* was launched. Now the parallel is completed with the publication of the *British Journal of Applied Physics*. Such a close parallel is indeed an indication of the real place that exists in each country for a journal devoted to the interests of the applied physicist. Significantly journals of almost the same name and purpose have recently been established in all of the principal countries of Europe.

We want in this brief note to congratulate Dr. Lang on this new venture of the Institute of Physics and join Dr. Toy in requesting "our many friends all over the world to support this journal by contributing to its pages, by subscribing to it, and by encouraging others to do so also."

ELMER HUTCHISSON

Microwave Angle Separation on a Two and One-Half Mile Overwater Path*

A. W. Straiton and A. H. LaGrone
The University of Texas, Austin, Texas
(Received August 29, 1949)

This paper describes the results of radio angle-separation measurements on a 2.50-mile path over Lake Buchanan near Austin, Texas. The measurements were made to test the angle-separation equipment developed at The University of Texas, and to compare these results with measurements obtained by rocking a 20-foot parabolic antenna.

I. INTRODUCTION

THE measurement of the angle-of-arrival of radio waves at near grazing angles is complicated by reflections from the ground. This paper describes measurements made to separate the direct wave from the ground-reflected wave, using a wave-length of 3.2 centimeters.

The measurements made in this report were for two rays arriving at angles separated by less than 0.5 degree. The magnitude of the reflected ray was measured to be 0.94 of the direct ray. Thus, these conditions of two rays of approximately equal magnitude arriving at such small angles of separation made the problem of their separation a very difficult one.

The angles measured by the phase-difference method are not singular, but are ambiguous. For a spacing of 5 feet and a wave-length of 3.2 centimeters, this ambiguity occurs every 1.2 degrees. This difficulty could be resolved by coarser methods of measurement.

Three different methods were used in the angle determinations.

1. The first method requires as data the relative phase and the relative signal strength of three vertically spaced antennas. This method will be referred to throughout this report as the "three-antenna method."

2. The second method requires as data the relative phase between two antennas, the signal strength at the two antennas relative to the maximum signal from a height-gain curve, and the minimum signal from a height-gain curve. This method will be referred to throughout this report as the "two-antenna method."

3. The third method requires as data the record of the signal received by a large antenna as its axis is tilted through an angle large enough to include the two-wave components. This method will be referred to throughout this report as the "amplitude method."

II. THEORY OF MEASUREMENT METHODS

1. Three-Antenna Method

This method assumes two plane or two spherical waves arriving at three vertically spaced antennas. If the resultant phase differences between the three antennas and the resultant relative signal strengths at the three antennas are known, the angles-of-arrival of each of the two-wave components, their relative magnitudes and their relative phase at the center antenna may be determined by calculations and charts. This mathematical relationship has been described elsewhere.[1,2]

Equipment has been built at the Electrical Engineering Research Laboratory of The University of Texas for obtaining these desired data. The basic phase-difference equipment has been described in Report 8,[3] and the operation characteristics of the three-antenna equipment has been described in Report 24.[4]

This method does not require moving the antenna system and hence gives instantaneous data.

2. Two-Antenna Method

The two-antenna method requires a height run over a wide enough range to include at least one signal strength minimum and one maximum. In addition, the phase difference between two points and the signal strength of these points relative to the signal strength maximum must be known.

For the five-foot spacing between antennas used in this work, the angle-of-arrival of the direct wave, θ_D, and the angle-of-arrival of the reflected wave, θ_R, are given by

$$\theta_D = -(1/300)(d\beta + A_z - A_0), \tag{1}$$

$$\theta_R = \theta_D - (1/300)(T_z - T_0). \tag{2}$$

In these equations, $d\beta$ is the phase of the upper antenna signal minus the phase of the lower antenna signal. A_z and T_z are terms which depend on the signal at the upper antenna relative to the maximum signal strength. A_0 and T_0 are the corresponding terms for the lower antenna. If the signal strength is increasing with height at the point in question, A is positive and T is negative. If the signal strength is decreasing with height, A is negative and T is positive.

The derivation of the basic relationships and the

* This work was sponsored by ONR, Contract N5ori-136, Project Order I.

[1] Hamlin, Seay, and Gordon, "New solutions to the problem of vertical angle-of-arrival of radio waves," J. App. Phys. **20**, 248 (1949).

[2] A. H. LaGrone and J. A. Hawkins, "Application of angle-of-arrival to two spherical waves," Memorandum No. 5, Electrical Engineering Research Laboratory, The University of Texas, 1948.

[3] F. E. Brooks, Jr. *et al.*, "Operating manual for phase difference equipment," Report No. 8, Electrical Engineering Research Laboratory, The University of Texas, 1948.

[4] A. W. Straiton *et al.*, "Operational Characteristics of angle separation equipment," Report No. 24, Electrical Engineering Research Laboratory, The University of Texas, 1949.

charts for obtaining A and T are given elsewhere.[5] The minus sign in front of the right-hand expression of Eq. (1) was omitted in reference 5.

3. Amplitude Method

This method involves the use of a large antenna which is tilted so as to determine its scanning pattern in a vertical plane. The antenna used was a 20-foot parabolic antenna with a rectangular face. This antenna was used by the Bell Telephone Laboratories in their early microwave angle-of-arrival measurements,[6] and was subsequently made available to the Electrical Engineering Research Laboratory. It is assumed that maxima in the scanning pattern occur when the tilt is such that the axis of the parabola is pointed in the direction of an incoming wave component.

III. MEASUREMENT SITE

1. Location

The radio path chosen was across Lake Buchanan located about 65 miles from Austin. The length of the path was about two and one-half miles. This length of path was chosen because (1) it made possible a range of angles up to about 0.5 degree; (2) it was long enough to reduce the effect of wave curvature; and (3) it was short enough to make the effect of meteorological changes small.

2. Transmitter

The transmitter was located on a 50-foot tower approximately 20 feet from the water's edge, and had a minimum height of 11 feet above the water. The transmitter assembly used a 2K39 Klystron with a c.w. output of 200 milliwatts.

3. Receivers

The 60-foot receiving tower, with the three receiving antennas mounted on an elevator, was located about 300 feet from the water's edge. The minimum height of the center antenna was 22 feet above the lake level. The vertical spacing of the three-antenna array was five feet. A reference transmitter was moved in front of the receiving antenna and surveyed into position when a reference angle was needed.

The Bell Laboratories' antenna was located near the 60-foot receiving tower and about 250 feet from the water's edge. A synchronous motor was provided for rocking the antenna at a uniform rate. However, in taking the scanning patterns used in this report, the antenna was tilted by hand in order to obtain its position more accurately. The center of the 20-foot parabola was 30 feet above the water.

[5] Straiton, Gordon, and LaGrone, "A method of determining the angle-of-arrival," J. App. Phys. **19**, 524–533 (1948).

[6] W. M. Sharpless, "Measurement of the angle-of-arrival of microwaves," Proc. Inst. Radio Eng. **34**, 837–845 (1946).

4. Meteorological Station

In order to compensate for changes in the meteorological conditions near the ground, a meteorological station was set up near the receiving site. A tilted 60-foot mast imbedded at the water's edge permitted raising a recording psychrometer a short distance out from shore in the direction of the prevailing wind. An anemometer was located near by to make a record of the wind speed and direction.

5. Angles-of-Arrival under Standard Conditions

The angular separation of the direct and reflected rays is a function of the height of the transmitter and the path length for standard meteorological conditions. This angular separation is shown for various transmitter heights in Fig. 1.

The angles-of-arrival of the direct and reflected rays are functions of transmitter and receiver height and path length. These angles-of-arrival as a function of transmitter height are plotted in Fig. 1 for a number of receiver heights with correction applied for the refraction associated with standard meteorological conditions.

IV. RESULTS OF MEASUREMENTS

1. Three-Antenna Method

Measurements were made of the angle-of-arrival by the three-antenna method for 164 cases. These were made for various combinations of transmitter and re-

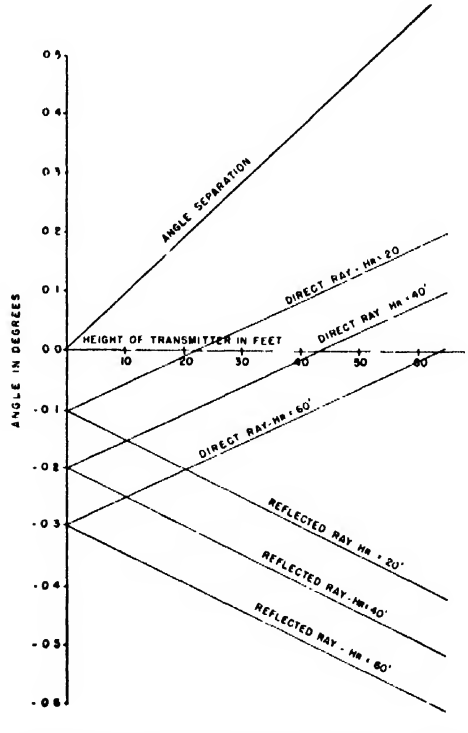


FIG. 1. Angles as a function of transmitter and receiver height.

TABLE I. Errors in angles-of-arrival of the direct and reflected rays.* Three-antenna method.

Transmitter height	Separation angle	Direct ray $.00° < \alpha \leq .02°$		$.02° < \alpha \leq .06°$		$.06° < \alpha \leq .10°$		$.10° < \alpha$		Reflected ray $.00° < \rho \leq .02°$		$.02° < \rho \leq .06°$		$.06° < \rho < .10°$		$.10° < \rho$	
11 feet	0.10°	11	(2)	33	(18)	8	(4)	2	(2)	3	(3)	5	(2)	4		42	(30)
21 feet	0.19°	2		1		1		3		2		1		3		1	
33 feet	0.29°	14	(7)	18		4		18	(16)	11	(2)	15	(3)	8	(3)	20	(15)
43 feet	0.38°	0		1	(1)	0		1	(1)	1	(1)	0		0		1	(1)
53 feet	0.47°	8	(1)	21	(5)	3	(2)	15	(14)	6	(2)	22	(4)	7		12	(11)
		21.3%	(35.2%)	45.1%	(47.3%)	9.8%	(11.0%)	23.8%	(5.6%)	14.0%	(18.3%)	26.2%	(41.5%)	13.4%	(17.1%)	46.3%	(23.2%)

* Numbers in parentheses are the number of cases which were rejected. Percentages in parentheses are based on the number of accepted cases.

ceiver heights. More cases were considered with the receiver at 30 and 35 feet than at any other height, since this was approximately the height of the center of the 20-foot antenna. More cases were considered with the transmitter at 11 feet than any other height, since this presented the smallest angular separation of the direct and reflected rays.

From the phase differences and signal-strength ratios between the three antennas, the angles-of-arrival of the direct and reflected rays were calculated by the method of reference 1.

The modified index-of-refraction profile was determined from meteorological measurements made simultaneously with the radio measurements. Since the

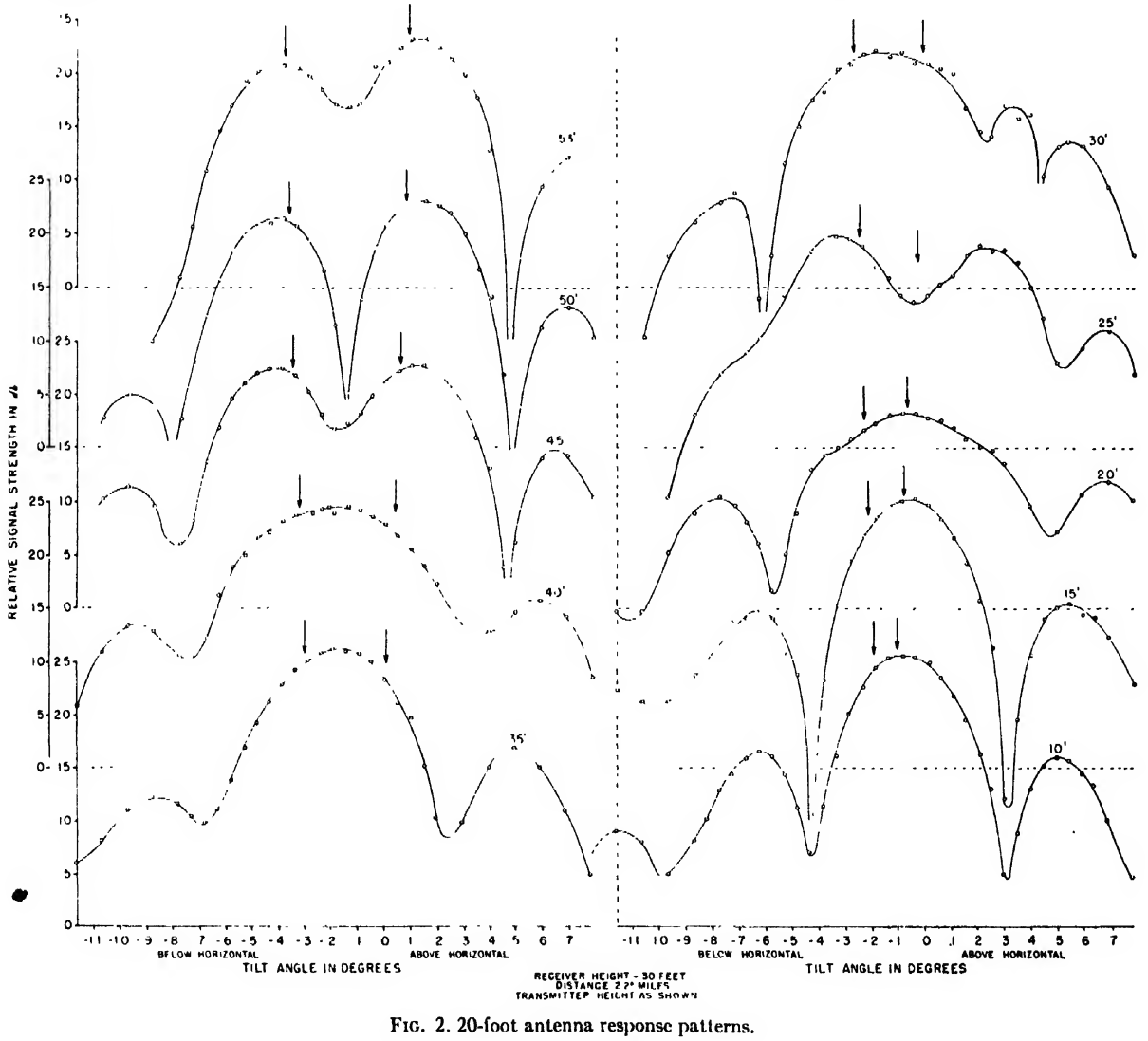


FIG. 2. 20-foot antenna response patterns.

TABLE II. Errors in angles-of-arrival of the direct and reflected rays.* Two-antenna method.

Transmitter height	Separation angle	Direct ray $.00° < \alpha \leq .02°$	$.02° < \alpha \leq .06°$	$.06° < \alpha \leq .10°$	$.10° < \alpha$	Reflected ray $.0° < \rho \leq .02°$	$.02° < \rho \leq .06°$	$.06° < \rho \leq .10°$
11 feet	0.10°	19	12	1	0	16	15	1	0
21 feet	0.19°	2	2	1 (1)	1 (1)	1	2 (1)	2	1 (1)
33 feet	0.29°	14 (2)	13 (2)	5	0	15 (1)	11 (1)	6 (2)	0
43 feet	0.38°	0	2 (1)	0	2 (2)	1	1 (1)	0	2 (2)
53 feet	0.47°	13	11 (2)	10 (5)	11 (11)	7	22 (5)	5 (3)	10 (10)
		40.7% (50.5%)	33.9% (38.5%)	13.6% (11.0%)	11.9% (0 %)	33.9% (42.9%)	43.2% (47.3%)	11.9% (9.9%)	11.0% (0 %)

* Numbers in parentheses are the number of cases which were rejected. Percentages in parentheses are based on the number of accepted cases.

changes in refraction with meteorological changes were small, linear values for index gradients were assumed for the height-range of the direct and of the reflected rays. The angles of Fig. 1 were corrected by use of the appropriate M-gradient, and the corrected values compared to those measured.

For a qualitative evaluation of the test, the results were grouped into four classes. The first was the group for which the radio measurements differed from the geometrical angles corrected for refraction by 0.02 degree or less. The second group included measurements for which the error was 0.06 degree or less, but greater than 0.02 degree. The third group included those with errors of 0.1 degree or less, but greater than 0.06 degree. The fourth group included those cases for which the error was greater than 0.1 degree. The groups might be classified as (1) extremely accurate, (2) acceptable, (3) inaccurate, and (4) useless.

The results of the 164 cases measured by the three-antenna method are shown in Table I. The division of the measurements between the four groups is shown in percentage in two ways at the foot of Table I. The first percentages are based on all of the measurements made. The second percentage excludes the cases for which the data would indicate that correct results could not be obtained. The number of cases omitted in the second percentage is shown in parentheses in the tabulation.

The excluded cases included several for which the direct and reflected waves are nearly in-phase or 180 degrees out-of-phase at the center antenna. It has been shown previously[1] that the method used becomes very inaccurate for these conditions. These conditions are known from the data to exist when the phase difference between the upper and middle and lower and middle antennas are equal, and the corresponding signal-strength ratios are equal.

All cases for which the ratio of the signal at two antennas exceeded 12 db were aslo excluded since the equipment was known to be inaccurate under these conditions. For a single-component wave arriving at the receiver, the signal-strength ratios will both be one and the phase difference between antenna pairs will be equal to each other. Under these conditions, the solution provided by the three-antenna method will be the true angle-of-arrival of the single-component wave plus a fictitious angle 0.3 degree above the true angle. In the case of the measurements made with the transmitter at 11 feet, the signal strength ratios are nearly one, and the angle between antenna pairs nearly equal to each other. This means that an extremely high degree of accuracy is required in obtaining the data if the angle-of-arrival of each component is to be determined. Apparently, this accuracy was not always met, and the result was that the apparatus measured only a single ray coming from approximately halfway between the transmitter and its reflection. The second angle did not exist but was introduced in the mathematical processes. This accounts for the fact that so many of the direct-ray measurement errors were between 0.02 and 0.06 degree, and the "reflected" ray errors were greater than 0.1 degree when the transmitter was at 11 feet.

The calculation of the angles-of-arrival by the three-antenna method requires subtracting the products of the signal-strength ratios and sines and cosines of the phase differences. When the products to be subtracted are of approximately the same magnitude, accuracy with which the difference is known is very low. For this reason, a very high degree of accuracy in the original data must be obtained. Cases where the difference of the two products was less than 0.1 were placed in the excluded group.

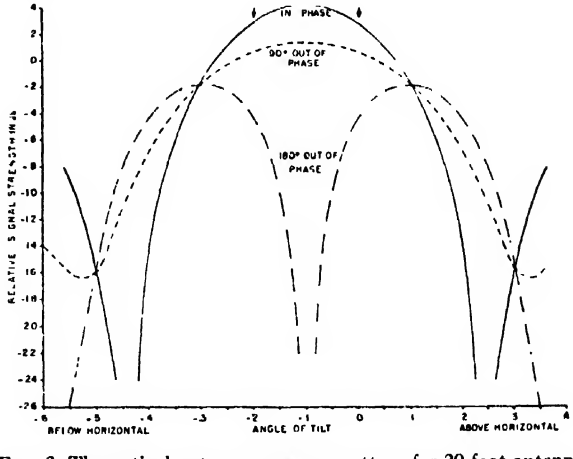


FIG. 3. Theoretical antenna response pattern for 20-foot antenna

TABLE III. Errors in angles-of-arrival of the direct and reflected rays.* Amplitude method.

Transmitter height	Separation angle	Direct ray $.00° < \alpha \leq .02°$	$.02° < \alpha \leq .06°$	$.06° < \alpha \leq .10°$	$.10° < \alpha$	Reflected ray $.00° < \rho \leq .02°$	$.02° < \rho \leq .06°$	$.06° < \rho \leq .10°$	$.10° < \rho$
11 feet	0.10°	0	0	0	1	0	1	0	0
21 feet	0.19°	0	0	1	2	1	1	0	1
33 feet	0.29°	0	0	0	2	0	0	0	2
43 feet	0.38°	0	0	1	1	0	0	1	1
53 feet	0.47°	1	1	0	0	0	0	2	0
		10%	10%	20%	60%	10%	20%	30%	40%

* Numbers in parentheses are the number of cases which were rejected. Percentages in parentheses are based on the number of accepted cases.

Although the results of this initial field test of the three-antenna method cannot be classed as highly satisfactory, it is felt that with improvements in the accuracy of the original data will come improvements in the accuracy of measuring the angles.

2. Two-Antenna Method

The two-antenna method required as data a height-gain curve and the phase difference between two points. The phase measurements of the three-antenna method provided the necessary phase difference data for the two-antenna method, and a height run after most of the three-antenna tests provided the necessary signal strength data.

One hundred and eighteen sets of measurements were made by the two-antenna method, and the results were classified in the same manner as for the three-antenna method. This classification is shown in Table II. Two sets of percentages are again shown: one for all the data, and one excluding the data for which the signal ratio between the antennas exceeded 12 db.

The height-gain curves agree precisely with those which would be obtained from two waves with a relative magnitude of 0.94. This equivalent reflection coefficient was virtually the same for all days, and was independent of transmitter and receiver heights.

The accuracy of the angle-of-arrival calculations by the two-antenna method is very satisfactory. However, this method has obviously disadvantages which limit its usefulness. A height-gain curve including signal strength maximum and minimum is not always obtainable. In addition, all of the needed data are not available instantaneously.

For fluctuating or changing conditions, this makes the angle determination difficult or impossible. However, the conditions of the present test made possible highly accurate evaluation of the angles-of-arrival by the two-antenna method.

3. Amplitude Method

Scanning patterns for the 20-foot antenna were made for a number of transmitter heights on a day when the atmospheric conditions were known to be nearly standard. These antenna patterns for five-foot intervals of the transmitter are shown in Fig. 2. The arrow in the diagram indicates the angle-of-arrival determined from geometry, corrected for refraction.

The angles-of-arrival as determined from the signal strength maxima are classified in Table III in the same way as for the previous methods.

As shown in a previous report,[7] the scanning patterns for two waves arriving simultaneously is greatly affected by the phase relationship of the two waves. This is illustrated by the theoretical scanning patterns shown in Fig. 3. When the two waves of nearly equal amplitude arrive in phase at the center of the antenna face, a single maximum results, positioned between the true angles-of-arrival. When the two arriving waves are 180 degrees out-of-phase at the center of the two antennas, two distinct maxima appear. However, these maxima indicate an angular separation of the two rays which is too large.

V. SUMMARY

The field test of the three-antenna angle-separating equipment showed that under the most desirable circumstances, the angles-of-arrival of two-wave components may be measured accurately. However, more accuracy is required in the original data if reliable results are to be obtained under all conditions.

The three-antenna system is basically a very significant one because it involves no time delay in taking the data. Because of this instantaneous action of the method, it is felt that further development is justified. Greater accuracy could be obtained by increasing the antenna spacing. This would, however, reduce the angle of ambiguity.

The two-antenna method showed a very high degree of accuracy and gave by far the best results of the three methods considered. However, the requirement of a height-run to determine the signal-strength maximum and minimum limits its usefulness.

The amplitude method appears to be unsatisfactory for separating two rays of nearly equal magnitude with angular separations comparable to the width of the basic antenna pattern. However, the difficulty of ambiguity of the measurement is eliminated in this method.

[7] LaGrone, Hamlin, and Straiton, "The indicated angle-of-arrival by phase front analysis," Report No. 12, Electrical Engineering Research Laboratory, The University of Texas, 1948.

The scanning patterns would not be reliable in a rapidly fluctuating field. The original experiments with the 20-foot antenna did not require the separation of two-wave components so nearly equal or separated by as small an angle as in the test described in this report.

VI. ACKNOWLEDGMENT

The Electrical Engineering Research Laboratory wishes to acknowledge its gratitude to the Bell Telephone Laboratories for the use of the parabolic antenna which provided the comparison test.

Immobility Phenomena and Reverse Driving Phenomena of the Electric Arc

SAKAE YAMAMURA

The Electrical Department, Faculty of Engineering, University of Tokyo, Tokyo, Japan

(Received June 22, 1949)

When an electric arc is driven by a magnetic field, under special conditions, curious phenomena are observed. An electric arc cannot be driven by the magnetic field when it is established at very small gaps between the electrodes. Neither can it be driven by the magnetic field for the instant just after the contacts have been broken. These phenomena are called immobility phenomena. In the present work, the immobility time and the driving velocity of the electric arc were measured under various air pressures. As the air pressure decreases, the immobility time increases and the driving velocity decreases. Under a certain low air pressure, the arc cannot be driven by the magnetic field. When the air pressure is lower than this critical value, the arc can be driven again by the magnetic field; but in this case, the arc moves in a direction that is opposite to the direction of the electromagnetic force. This phenomenon is called the reverse driving phenomenon. The causes for these phenomena are considered.

I. INTRODUCTION

WHEN an electric arc is established in the presence of the magnetic field, curious phenomena are observed. Under normal conditions, the arc moves in the same direction as the electromagnetic force. Under some special conditions, such as at a very small gap,[1] just after the contacts have been broken,[1,2] and at low air pressures,[3,4] the arc cannot be driven by the magnetic field. Scars on the electrodes made the electric arc remain stationary. Under still other conditions, the arc moves in a direction that is opposite to the direction of the electromagnetic force.

These phenomena were observed independently by several persons including the author.[1,3] But interrelations between these phenomena have been missed. Some of the experimental results and considerations about these phenomena will be presented in the following reports.

II. EXPERIMENTAL APPARATUS

A picture of the experimental apparatus that was used is shown in Fig. 1. The electric arc is established by separating two parallel copper electrodes. The electrodes are placed horizontally and are sealed in a glass vessel which is connected to a rotary vacuum pump. The driving magnetic field is applied perpendicularly to the arc as is shown by the arrow H in Fig. 1. The electrodes are pulled apart and shorted together by an electromagnet (there is no relation between this magnet and the driving magnetic field). The magnet which pulls apart and shorts together the electrodes is controlled by a rotary switch that is connected to the shaft of a rotating mirror. The image of the arc is focused through a lens and is reflected on the screen by a rotating mirror. Since the mirror is rotating, the image of the arc will repeatedly sweep from the bottom to the top of the screen. The rotating mirror revolves at a constant velocity. In this way, the loci of the images of several successive electric arcs will appear in the same position on the screen. If the arc is stationary, the locus of its image on the screen is a vertical line (as from a to b). If the arc is driven by the magnetic field, the locus of its image is an oblique line (as from b to c). The horizontal velocity of the arc can be determined by measuring the angle θ of the locus on the screen.

III. OBSERVATIONS

1. The Magnetic Driving of a Contact-Arc

An electric arc which is generated by breaking contacts is named a contact-arc. The movements of a contact-arc which was driven by the magnetic field were observed with the apparatus described in Section II. The locus of the image of the contact-arc was the same as the locus shown on the screen in Fig. 1. This locus indicates that the contact-arc is stationary for an instant just after the contacts have been broken; then it begins suddenly to be driven by the magnetic field. After being driven for a moment, the arc goes out. This phenomenon was named the immobility phenomenon.[1,2] The author proposes to name it the momentary immobility phenomenon to distinguish it from the permanent immobility phenomenon. The permanent immobility phenomenon will be discussed later.

[1] S. Yamamura, J. Inst. Elec. Eng. Japan **61**, 660 (1941).
[2] S. Fukuda, J. Inst. Elec. Eng. Japan **62**, 161 (1942).
[3] S. Yamamura, Report of Annual Meeting of Inst. Elec. Eng. Japan, 16 (1948).
[4] G. J. Himler and G. L. Cohn, Elec. Eng. **67**, 1148 (1948).

The immobility time and the driving velocity of the contact-arc were measured under various air pressures. Figure 2 shows one example of these measurements.

As the air pressure decreases, the driving velocity decreases and the immobility time increases. At a certain pressure (in Fig. 2, 350 mm Hg), the electric arc cannot be driven by the magnetic field, and it remains permanently stationary. This phenomenon is called the permanent immobility phenomenon.

When the pressure is reduced below this critical value (350 mm Hg), the arc can be driven again. But

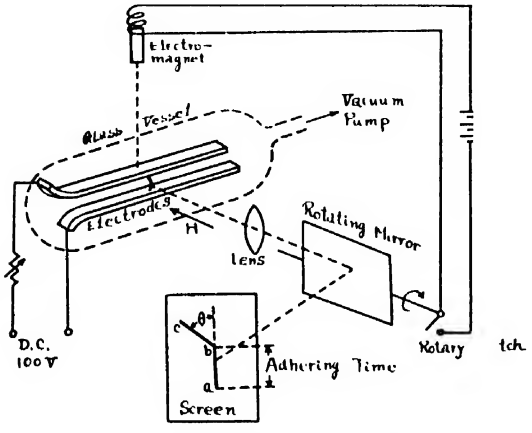


FIG. 1. Apparatus for experimenting with the magnetic driving of the electric arc.

in this case, the arc moves in a direction that is opposite to the direction of the electromagnetic force. This phenomenon is called the reverse driving phenomenon.

As the air pressure decreases further, the reverse driving velocity increases greatly and the immobility time decreases to almost a zero value, as one can determine from Fig. 2.

The same sequence of phenomena occurs when the arc current and the driving magnetic field are of different values; but, in these cases, the critical air pressure at which permanent immobility occurs is different. These relations are shown in Fig. 3. The critical air pressure (the pressure at which permanent immobility occurs) increases as the magnetic field intensity increases. This is contrary to the statements of Himler and Cohn,[4] who reported that the critical air pressure decreased as the magnetic field increased.

The same experiments were performed with several different electrode materials to decide whether they would show permanent immobility and reverse driving. One example using aluminum electrodes is shown in Fig. 4. The aluminum electrodes have a wider pressure range where permanent immobility occurs.

All of the experimental results are summarized in Table I. The materials are arranged as grouped by Druyvesteyn and Penning.[5] These authors arranged the various metals into nine groups according to the mobility of the cathode spot of the metal.

Permanent immobility was observed with all of the materials that were examined. When the air pressure was reduced to 50 mm Hg, reverse driving was observed in all of the groups except groups I and II.

2. Some Other Immobility Phenomena of the Electric Arc

When an electric arc is established at a very small gap between the two parallel electrodes, it cannot be driven by the magnetic field. In this case, even under a normal atmospheric pressure, permanent immobility occurs.

Since a molten metal bridge exists between the electrodes which are separated very slowly, the arc does not appear until the electrodes have been separated to the displacement (D_1) which is shown by the broken line in Fig. 5. The continuous line of Fig. 5 indicates the displacement (D_2) of the movable electrode at which the arc begins to move quickly. Between displacements (D_1) and (D_2) the arc cannot be driven by the magnetic field.

Scars on the electrodes, especially on the cathode, also make the electric arc remain stationary.

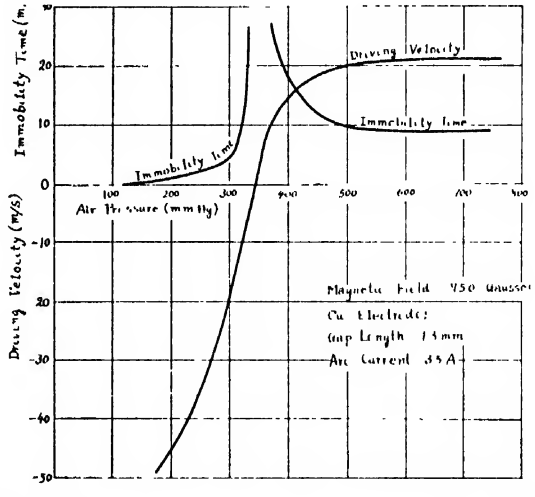


FIG. 2. The relations between the air pressure, the immobility time and the driving velocity of the contact-arc.

IV. CONSIDERATIONS

C. G. Smith[6] reported about the reverse driving of the cathode spot of a mercury arc. He explained this phenomenon by using the negative Righi-Leduc effect, but his explanation does not seem plausible. Not only do the metals with a negative coefficient of the Righi-Leduc effect (for instance Mg, Cu and Ag), but also the metals with a positive coefficient (for instance Fe) show reverse driving. It is impossible that the coeffi-

[5] M. J. Druyvesteyn and F. M. Penning, Rev. Mod. Phys. **12**, 87 (1940).

[6] C. G. Smith, Phys. Rev. **62**, 48 (1942).

TABLE I. Classification of the electrode materials according to the characteristics of the arc on them.

	I	II	III	IV	V	VI	VII	VIII	IX
Electrode material	pure C	pure solid Zn pure solid Pb	normal Cu normal Fe normal Ag	Al Sn	impure Fe	very pure Cu very pure Fe very pure Ag	liquid Hg*	normal Cu	(Al) (Sn)
Impurities present	no	no	yes	yes	yes	no	no		
Permanent immobility	yes	yes	yes	yes	yes	yes	yes	yes	yes
Reverse driving	no	no	yes	yes	yes	yes	yes	yes	yes

* According to C. G. Smith (see reference 6).

cient of the Righi-Leduc effect changes its sign from positive to negative with any change of air pressure.

Our explanation involves the electron diffusion theory. Himler and Cohn[4] and the author[3] applied almost the same explanation. An enlarged picture of a cathode spot is shown in Fig. 6. When the magnetic field is applied perpendicularly from the front to the back of the paper, the loci of electrons leaving the cathode are bent to the right side by the magnetic field as is shown in Fig. 6. Thus there ensues the rightward movement of the arc, which is the normal magnetic driving of the arc.

When the atmospheric pressure is much lower, that is, when the mean free path of the electrons is longer, the path of the electrons appears to be influenced to a great degree by the magnetic field. Many electrons with energy corresponding to the cathode fall, diffuse outwards in the positive space charge cloud. The loci of these electrons are bent by the transversal magnetic field as is shown in Fig. 6. The electrons diffusing to the left side are accelerated by the electric field, and thus these electrons efficiently ionize the gas molecules by collision. The electrons diffusing to the right side are decelerated by the electric field because these electrons are moving in an opposite direction to the electric field. Thus the electrons moving to the right side cannot efficiently ionize the gas molecules by collision. In this way, the positive space charge cloud spreads to the left side.

If this leftward spreading of the positive space charge cloud overcomes the rightward spreading (the normal magnetic driving of the arc), then reverse driving ensues, If these two spreadings balance each other, then permanent immobility ensues. These phenomena can occur only when the diffusing electrons have a comparatively long mean free path, and when their loci are bent notably by the magnetic field. Let us check this.

From Fig. 3 we can see that the critical air pressure at which permanent immobility occurs is 100–200 mm Hg at 360 gausses. The electron mean free path (λ_0) in the atmosphere is 3.6×10^{-5} cm at 0°C. The temperature of the positive column of the arc is 5000–10,000°K. A great many electrons in the positive space charge cloud, which have been accelerated by the cathode fall, have a very high temperature. Since this is true, the temperature of the gas molecules in the positive space charge cloud would be higher than the temperature of the positive column. Therefore the temperature of the gas molecules could be assumed to be 10,000°K. Then the electron mean free path (λ_r) at 100 mm Hg is

$$\lambda_r=\lambda_0\times\frac{760}{p}\times\frac{T}{273}=3.6\times10^{-5}\times\frac{760}{100}\times\frac{10{,}000}{273}=1.0\times10^{-2}\text{ cm}. \quad (1)$$

The radius R of a circle, which an electron with the energy of V electron-volts follows under the influence of the perpendicular magnetic field of H gausses, is

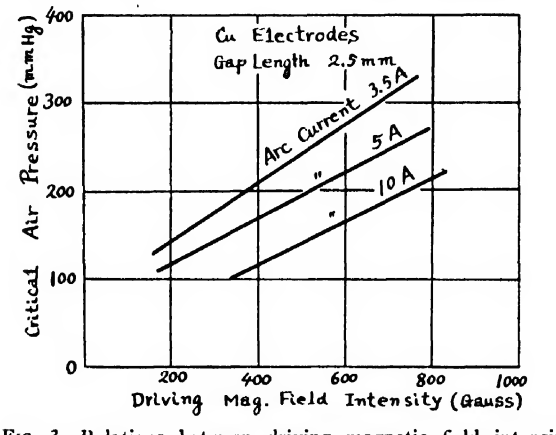


FIG. 3. Relations between driving magnetic field intensity and critical air pressure at which the permanent immobility occurs.

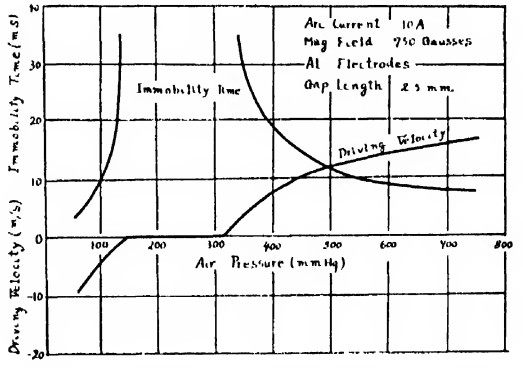


FIG. 4. The same relations as in Fig. 2. But the electrodes are made of aluminum.

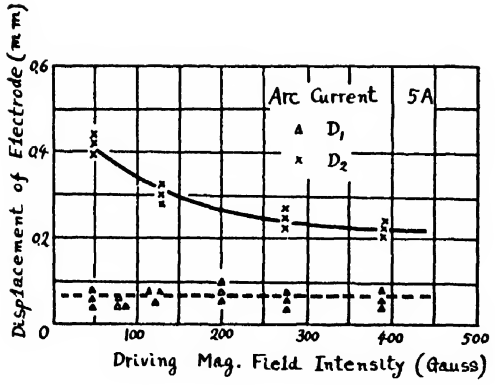


FIG. 5. Relations between driving magnetic field intensity, and electrode displacement D_1 at which an electric arc appears, and electrode displacement D_2 at which an electric arc begins to be driven. The electrodes are clean.

given by the formula

$$R=\frac{3.4(V)^{\frac{1}{2}}}{H}\ \text{cm}. \tag{2}$$

Many of the electrons diffusing outward in the positive space charge cloud have been accelerated by the cathode fall before they make an elastic collision with a gas molecule. The energy of the diffusing electrons (the electrons that have left the cathode fall) may be assumed to be 8 electron-volts. This corresponds to the cathode fall of the copper electrode. When H is 360 gausses, then

$$R=\frac{3.4(8)^{\frac{1}{2}}}{360}=2.7\times10^{-2}\ \text{cm}. \tag{3}$$

The electron mean free path and the radius are nearly equal. Thus it is probable that the loci of the many diffusing electrons are bent notably by the magnetic field as shown in Fig. 6. Although it is accompanied by some uncertainties due to complications in the cathode spot, this calculation would confirm the electron diffusion theory of reverse driving and of the immobility phenomena under a low air pressure.

As the driving magnetic field increases, the radius R of the circular loci of the electrons given by Eq. (2) becomes smaller. Thus permanent immobility will occur at a higher air pressure with an increase in the driving magnetic field.

Momentary immobility of the electric arc might be due to the fact that the temperature of the cathode spot might be very high when the arc is first established. Thus for an instant, the electron mean free path is relatively long and much curved by the magnetic field. Permanent immobility at very small gaps may be attributed to the same reason. That is, the gas temperature is very high.

Easily vaporizable metals (groups III, IV in Table I) have a relatively long time of momentary immobility. Very easily vaporizable metals (group II in Table I) have a very long time of momentary immobility and show no reverse driving. Materials with a thermoemissive cathode spot (group I in Table I) show no reverse driving.

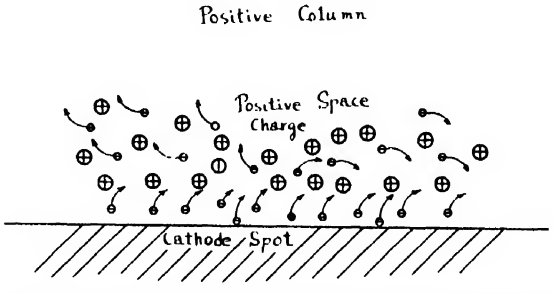


FIG. 6. Enlarged picture of the cathode spot. ⊕ positive ions. ⊖ electrons. The magnetic field is directed perpendicularly from the front to the back of the paper.

V. CONCLUSION

From the characteristics of the immobility phenomena and reverse driving phenomena which were reported, it is evident that they must be taken into consideration in the magne-blast circuit breaker. Momentary immobility not only would lengthen the interrupting time of the breaker, but it would also damage the most important parts of the contacts. Permanent immobility would make the breaker lose its interrupting ability at an altitude of several thousand meters. An air blast could protect the breaker from these phenomena effectively.

There remain many matters to prove experimentally. Further experiments and details of practical considerations mentioned will be reported later.

The author expresses his sincere thanks to Professors H. Yamashita, S. Fukuda, and S. Hoh for their guidance.

The author thanks sincerely Mr. J. P. LeGeros for his kind help with the manuscript. The author also thanks Mr. M. Yasui very much for his effective assistance.

Measurements of Mechanical Properties of Polyisobutylene at Audiofrequencies by a Twin Transducer*

ROBERT S. MARVIN,† EDWIN R. FITZGERALD,‡ AND JOHN D. FERRY
Departments of Chemistry and Physics, University of Wisconsin, Madison, Wisconsin
(Received July 25, 1949)

Apparatus has been developed for measuring the dynamic viscosity and rigidity of soft rubber-like solids in small oscillating deformations. A plate rigidly attached between two identical coils in two permanent magnets shears a pair of disk-shaped samples when a driving current is passed through one coil. The open circuit voltage from the other coil is compared in amplitude and phase with the driving current by a method in which all measurements are in the form of settings of a potential divider. The apparatus has two advantages over the more familiar resonance devices: (a) the amplitudes of motion, which need not be measured directly, are extremely small, thus minimizing any non-linear effects, or temperature change due to heat dissipation; (b) a continuous range of frequencies, spaced as closely as desired, is available without adjusting masses. The dynamic rigidity and viscosity of two samples of polyisobutylene, of molecular weights 1.2 and 0.47 million, have been measured at 15, 25, and 35°C at frequencies from 20 to 600 cycles/sec., with a reproducibility of five percent or better. The rigidity increases and the viscosity decreases with increasing frequency. Both decrease with increasing temperature, and are almost independent of molecular weight. The temperature dependence is quantitatively explained by the assumption that stresses relax by flow processes whose apparent activation energies are all identical with that characterizing the steady flow viscosity.

INTRODUCTION

THE interpretation of equilibrium rubber-like elasticity in terms of a three-dimensional network is well established.[1,2] In agreement with this theory, uncrosslinked linear polymers such as raw rubber and polyisobutylene have no equilibrium elasticity; when a sample is held at constant strain, the stress decays to zero.[3] Nevertheless, these unvulcanized rubbers show elasticity under dynamic or non-equilibrium conditions;[4] in a polyisobutylene of molecular weight about one million, the stress-strain ratio 12 sec. after application of stress[3] corresponds to an elasticity modulus of about 5×10^6 dyne/cm^2, and dynamic measurements at 100 cycles/sec. correspond to a dynamic Young's modulus of about 12×10^6 (see below), as compared with equilibrium values of the order of 10×10^6 for typical vulcanized rubbers. The origin of this non-equilibrium elasticity is not well understood. Knowledge of the phenomenon can be advanced by quantitative measurements of mechanical properties in small oscillating deformations.

Although many experimental methods for measuring dynamic properties of rubber-like solids have been described recently,[5–7] most are not very well adapted for study, over a range of audiofrequencies, of uncrosslinked materials which cannot support static stress even for relatively brief periods. For measurements on such soft polymers, as well as on swollen gels and highly concentrated solutions, we have constructed a twin electromagnetic transducer. This instrument has the advantages that the shear strains employed are very small (of the order of 10^{-5} to 10^{-3}), measurements can be made over a continuously variable frequency range (not being restricted to operation at mechanical resonance), and all the data are obtained in a particularly simple and precise form—as settings of a potential divider.

DESCRIPTION OF APPARATUS

Twin voice coils, wound on Lucite forms, are joined by short brass tubes to a brass plate, about one inch square (Fig. 1), and are suspended by eight piano wires, diameter 0.007 in., attached to small plates soldered to the tubes. The wires are adjusted under moderate tension so the coils are accurately centered in the air gaps of twin loudspeaker magnets (Jensen[8] Model PMB S3575). Identical disk-shaped samples can be strained in shear between the central brass plate and the ends of two fixed brass cylinders. The latter are held by a split steel block, also bolted to the base plate, and can be clamped by tightening two screws through the block, one of which is shown. Before clamping, the cylinders are advanced by calibrated lead screws, and the distance between cylinder and plate, representing the thickness of the sample, can be read on each side on circular scales.

A photograph of the apparatus with one magnet removed is shown in Fig. 2. The base plate rests in a box of sawdust on a concrete block mounted on cork. For temperature control, the apparatus is covered by a Celotex box connected by a rubber bellows to an air thermostat. The temperature of the central steel block, measured by a thermometer in an oil-filled well, is controlled to ±0.3°C or better. Tests with a thermocouple have shown that under the conditions of our experiments (which involve extremely

* Presented in part at a meeting of the American Physical Society, Madison, Wisconsin (June 21, 1948) and at a meeting of the Rubber Division of the American Chemical Society, Los Angeles (July 22, 1948).

† Carbide and Carbon Chemicals Corporation Fellow in Physical Chemistry (1947–49). Present address: Rubber Section, National Bureau of Standards, Washington 25, D. C.

‡ Department of Physics.

[1] Guth, James, and Mark, Adv. Colloid Sci. **2**, 253 (1946).

[2] P. J. Flory, Ind. Eng. Chem. **38**, 417 (1946).

[3] Andrews, Hofman-Bang, and Tobolsky, J. Polymer Sci. **3**, 669 (1948).

[4] Cf. K. H. Meyer and A. J. A. Van der Wyk, J. Polymer Sci. **1**, 49 (1946).

[5] A. W. Nolle, J. App. Phys. **19**, 753 (1948).

[6] J. H. Dillon and S. D. Gehman, Ind. Rubber World **115**, 16, 217 (1946).

[7] W. Kuhn and O. Künzle, Helv. Chim. Acta **30**, 839 (1947).

[8] Jensen Manufacturing Company, Chicago, Illinois.

small amplitudes of oscillation and hence very little heat dissipation) the temperature of the sample and the block are identical within 0.05°.

When an alternating e.m.f. is applied to one coil, the samples are subjected to sinusoidally varying shear strain, and an e.m.f. is generated in the second coil. To calculate the dynamic mechanical properties of the samples from transducer relationships, it is necessary to compare the applied and generated voltages in amplitude and phase. For this purpose, the electrical circuit shown in Fig. 3 is employed. The oscillator is a Hewlett-Packard[9] Model 200 CR. Four non-inductive resistors in series form a potential divider represented by R_1 and R_4, which is followed by a three-dial decade voltage divider R_2, General Radio[10] Type 654A, with a total impedance of 10,000 ohms. If we consider the input impedance at $VTVM$ as infinite (it is actually 0.5 megohm, and this approximation involves a maximum error of 0.5 percent) the potential E_1 appearing at the output of the decade box R_2 will be αRI_1, where α is the fractional setting of the decade box, $1/R = 1/R_1 + 1/R_2$, and I_1 is the current through the driving coil. The values of the resistors composing R_1 and R_3 were chosen so that R would be exactly 1, 10, or 100 ohms, plus a fourth resistance of about 1000 ohms to give the proper load across the oscillator. The potential E_1 is opposed to the e.m.f. of the driven coil, $\mathbf{E}_2$, and the resultant $\mathbf{E}_r$ is read on a Ballantine[11] Model 302 vacuum tube voltmeter, preceded by a Model 220 amplifier.[11] The high input impedance of the amplifier insures that no appreciable current is drawn from the driven coil. By means of a switch, E_1 may also be read separately. All connections are carefully shielded.

THEORY AND CALCULATIONS

From the principles of electromagnetic transducers[12] it can be readily shown that

$$\mathbf{E}_2 = (B_1 l_1 B_2 l_2 \times 10^{-9}/\mathbf{Z_M}) I_1, \tag{1}$$

where B_1 and B_2 are the magnetic field strengths of the speakers, l_1 and l_2 the lengths of wires in the coils, and $\mathbf{Z_M}$ the mechanical impedance defined as the ratio of the driving force on coil 1 to the velocity of coil 2. The real part of $\mathbf{E}_2$ is taken in phase with I_1. The voltmeter provides only the absolute value, $|\mathbf{E}_2|$. However, its components can be separated by opposing it to E_1 (which is in phase with I_1 and therefore written as real) as described and using the following procedure.

Denoting $B_1 l_1 B_2 l_2 \times 10^{-9}$ by K^2, and using the admittance $\mathbf{Y_M} = 1/\mathbf{Z_M} = Y_M' - iY_M''$, we have $\mathbf{E}_2 = K^2(Y_M' - iY_M'')I_1$, and the resultant voltage $\mathbf{E}_r$ is

$$\mathbf{E}_r = E_1 - \mathbf{E}_2 = [\alpha R - K^2 Y_M' + iK^2 Y_M'']I_1. \tag{2}$$

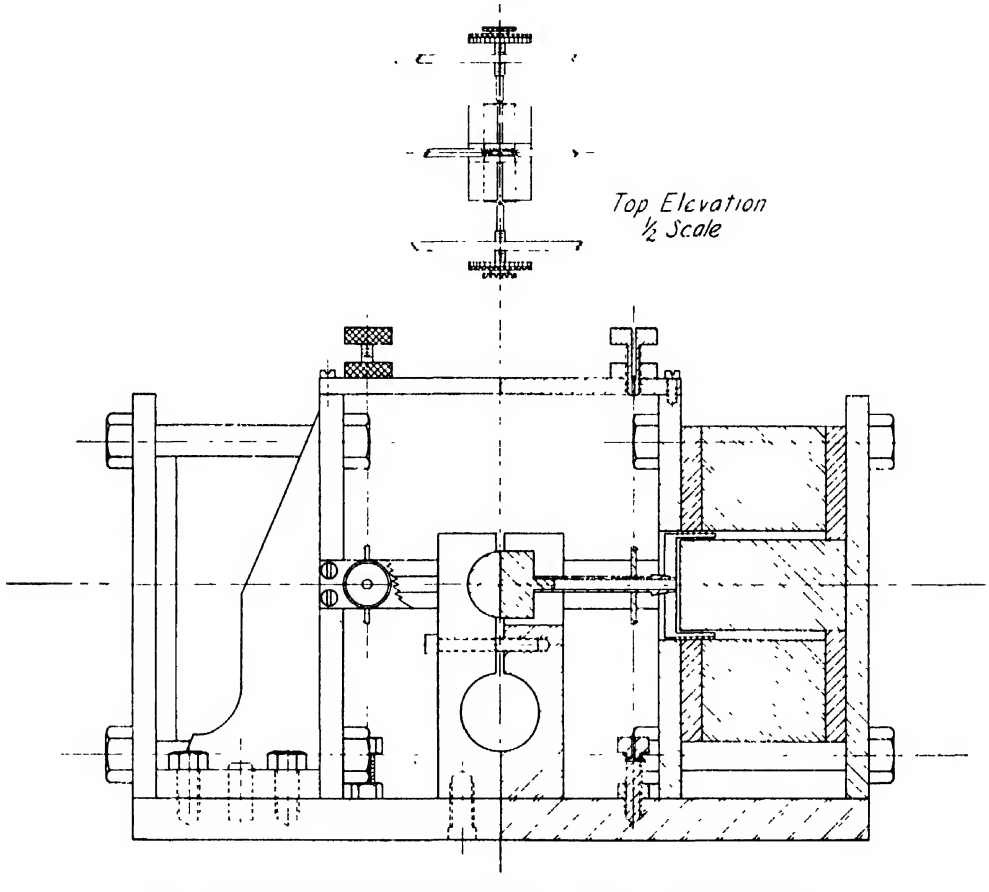


FIG. 1. Quarter-section drawing of twin transducer. Base plate is 12 in. long.

[9] Hewlett-Packard Company, Palo Alto, California.
[10] General Radio Company, Cambridge, Massachusetts.
[11] Ballantine Laboratories, Inc., Boonton, New Jersey.
[12] W. P. Mason, *Electromechanical Transducers and Wave Filters* (D. Van Nostrand and Company, Inc., New York, 1942).

When the potential divider setting α is varied at constant driving current and frequency, $|\mathbf{E}_r|$ goes through a minimum when the first two terms in the coefficient of I_1 cancel. At this point, the divider setting determines K^2Y_M' and the voltmeter reading determines K^2Y_M''; the coil e.m.f. $|\mathbf{E}_2|$ provides a check calculation:

$$K^2Y_M' = (\alpha R)_{\min}$$
$$K^2Y_M'' = |\mathbf{E}_r|_{\min}/I_1$$
$$K^2(Y_M'^2+Y_M''^2)^{\frac{1}{2}} = K^2|\mathbf{Y_M}| = |\mathbf{E}_2|/I_1. \qquad (3)$$

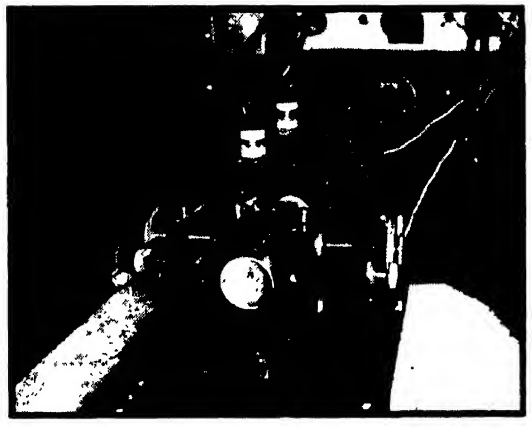


FIG. 2. Photograph of twin transducer with one magnet removed.

Since only ratios of voltage to current are required, the voltmeter scale readings for $|\mathbf{E}_r|$ and $|\mathbf{E}_2|$ are in practice matched by adjusting the potential divider with $\mathbf{E}_2$ switched out, at constant I_1, and the ratios $|\mathbf{E}_r|/I_1$ and $|\mathbf{E}_2|/I_1$ are thus obtained directly without relying upon calibration of the voltmeter scale. The check given by the last of Eqs. (3) is within one percent unless fluctuation in I_1 or frequency has occurred during the measurements.

The minimum is rather shallow; in some cases its location is facilitated by placing a resistance in series with the d.c. output of the vacuum tube voltmeter, and measuring the voltage drop across this resistance with an ordinary potentiometer and sensitive galvanometer.

The mechanical properties of the material subjected to sinusoidally varying shear are described by the dynamic rigidity, G' (real part of the complex dynamic shear modulus), and the dynamic viscosity, η' (real part of the complex dynamic viscosity), which have been defined elsewhere.[13,14] If the mass of the samples can be neglected, or their thickness is small compared to the wave-length of shear waves in the material, these quantities are related to the mechanical admittance as follows:

$$Y_M'/|\mathbf{Y_M}|^2 = R_M = R_M^0 + 2(A/h)\eta' \qquad (4)$$
$$Y_M''/|\mathbf{Y_M}|^2 = X_M = X_M^0 - 2(A/h)G'/\omega \qquad (5)$$

where R_M and R_M^0 are the mechanical resistances of the moving system with and without samples, X_M and X_M^0 are the corresponding mechanical reactances, A and h are the area and thickness of the two identical disk-shaped samples, and ω is the circular frequency. If the propagation of an incipient shear wave in the sample is taken into account, the contribution of the samples to the mechanical impedance is given by a hyperbolic cotangent function[15] which can be expanded to give a better approximation replacing Eq. (5):

$$Y_M''/|\mathbf{Y_M}|^2 = X_M = X_M^0 - 2(A/h)G'/\omega + 2\omega m/3 \qquad (6)$$

[13] Ferry, Sawyer, and Ashworth, J. Polymer Sci. **2**, 593 (1947); cf. H. Leaderman, J. Colloid Sci. 4, 193 (1949).

[14] Smith, Ferry, and Schremp, J. App. Phys. **20**, 144 (1949).

[15] R. S. Marvin, Ph.D. thesis, University of Wisconsin (1949).

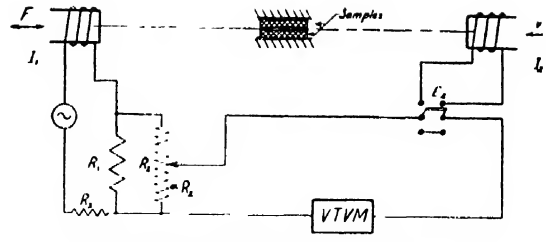


FIG. 3. Electrical circuit used with transducer.

where m is the mass of each sample. In practice, Eqs. (4) and (6) are employed, although the contribution of the term in m is slight (cf. Table I).

CALIBRATION

At each frequency, three apparatus constants are required: K^2, R_M^0, and X_M^0. The ratios R_M^0/K^2 and X_M^0/K^2 are readily determined from measurements without samples, using Eqs. (3)–(5). These values are plotted against the frequency in Fig. 4. Some isolated determinations showed that they were the same at 25° and 35°C.

The above discussion has tacitly assumed that the contributions of samples and apparatus to mechanical resistance and reactance are simply additive, with no interaction. If the coils, tubes, and plate moved as a rigid unit, this would be the case, and, moreover, X_M^0 would be equal to $\omega M - s_1/\omega$, where M is the mass of the coil-tube-plate assembly (68.2 g), and s_1 the elastance of the supporting wires. At frequencies well above the resonance $\omega_0 = (s_1/M)^{\frac{1}{2}}$, which corresponds to $\nu \simeq 24$ cycles/sec., X_M^0 should be simply ωM. Figure 4 shows that this is a good approximation up to about 200 cycles/sec.; however, at higher frequencies X_M^0/K^2 falls off from the $\omega M/K^2$ line, goes through a maximum, and falls to zero at about 3000 cycles/sec. (not shown). This behavior is attributed to a mode of vibration in which the coils move in and out symmetrically about a node at the plate, owing to compliance of the tubes, with a resonance frequency of about 3000 cycles/sec.[16] Because of this complication, no measurements of G' have been reported above 600 cycles/sec. with the present apparatus. Up to this frequency, the additivity expressed in Eqs. (4)–(6) has been assumed; its validity is shown by the agreement in G' and η' calculated from measurements on samples of different sizes (see below).

The constant K^2 is obtained by plotting $\omega Y_M''/|\mathbf{Y_M}|K^2 (= \omega X_M^0/K^2)$, determined from measurements without samples, against ω^2. If the coil-tube-plate assembly moves as a rigid unit, this should give a straight line with a slope of M/K^2 and an intercept of $-s_1/K^2$. From 20 to 100 cycles/sec. an excellent line is obtained from whose slope K^2 is found to be 8.5×10^4 ohm-dyne-sec.-cm^{-1}. This value is considered to be independent of frequency.[17, 17a]

[16] There are other indications that, even at fairly low frequencies, the mechanical impedance does not correspond to simple motion of a rigid rod, for which X_M would be $\omega M - S_M/\omega$, where $S_M = s_1 + 2(A/h)G'$. (1) There is a sharp discontinuity in the slope of the X_M^0/K^2 calibration curve at about 680 cycles (Fig. 5). (2) Values of G' obtained by applying the simple equation above to the value of the resonance frequency observed with a sample are about ten percent higher than those obtained from Eqs. (3) and (6). (3) If very thin samples are employed, two or even three resonances are observed below 1000 cycles, indicating strong interaction with the high frequency resonances. No such samples were used in obtaining the data reported here.

[17] A more elaborate analysis of calibration measurements at high frequencies, in which the compliance of the tubes is taken into account, shows that K^2 is independent of frequency at least up to 1500 cycles.

[17a] A recent method for the absolute calibration of electromechanical transducers [H. M. Trent, J. App. Mech. **70**, 49 (1948); S. P. Thompson, J. Acous. Soc. Am. **20**, 637 (1948)] has

TABLE I. Measurements and calculations of dynamic mechanical properties of polyisobutylene, $\bar{M}_\eta = 1.2\times10^6$, at 15°C; $m = 0.163$ g, $2A/h = 20.3$ cm.

Frequency, cycles/sec.	R_M/K^2 ohm^{-1}	$R_M{}^0/K^2$ ohm^{-1}	$\eta'\times10^{-3}$ poises	X_M/K^2 ohm^{-1}	$X_M{}^0/K^2$ ohm^{-1}	$2\omega m/3K^2$ ohm^{-1}	$G'\times10^{-6}$ dyne/cm^2
30	3.533	0.04	14.61	−4.932	0.055	0.000	3.94
45	3.069	0.04	12.68	−3.581	0.163	0.000	4.43
60	2.793	0.04	11.53	−2.840	0.258	0.000	4.89
90	2.424	0.05	9.94	−1.974	0.428	0.001	5.69
150	2.008	0.07	8.11	−1.022	0.741	0.001	6.96
200	1.812	0.08	7.25	−0.552	0.980	0.002	8.06
240	1.686	0.08	6.72	−0.211	1.160	0.002	8.66
300	1.514	0.08	6.00	+0.242	1.421	0.002	9.31
400	1.343	0.08	5.29	0.871	1.885	0.003	10.67
500	1.247	0.08	4.89	1.563	2.270	0.004	9.31
600	1.111	0.08	4.32	1.971	2.660	0.005	10.9
700	0.820	0.07	3.14	2.383	3.160	0.006	—
800	0.928	0.07	3.59	2.809	3.420	0.006	—
900	0.830	0.06	3.23	3.222	3.660	0.007	—
1200	0.660	−0.01	2.80	4.007	4.155	0.010	—
1500	0.349	−0.30	2.72	4.164	4.170	0.012	—

RESULTS ON POLYISOBUTYLENE

Two unfractionated samples of polyisobutylene were studied at 15°, 25°, and 35°C. The first (viscosity-average molecular weight 1.2 million[18]) was furnished through the kindness of Dr. John Rehner, Jr., Esso Laboratories. The second (viscosity-average molecular weight 0.47 million[18]), originally from Dr. Rehner's laboratory, was kindly given us by Dr. Rodney D. Andrews, Princeton University. Portions were compressed to sheets of uniform thickness (the relaxation of stress hastened by warming to 60°), and disks were cut by dies. In each case the disk was weighed and its volume was calculated, taking the density as 0.91. The volume divided by the square of the thickness after compression (determined by the cylinder lead screws in the apparatus, and often checked in place with a thickness gauge) gives the ratio A/h.

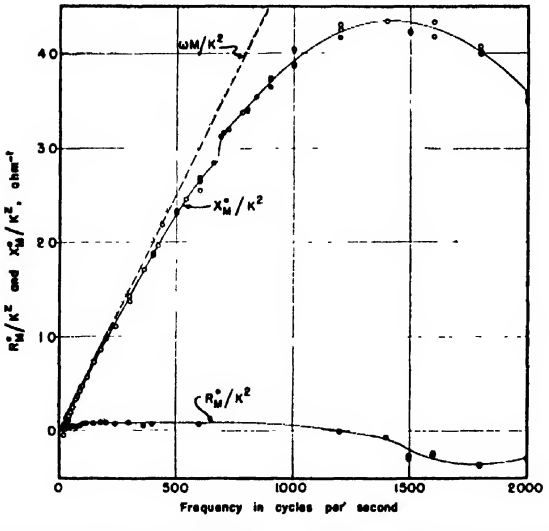


FIG. 4. $R_M{}^0/K^2$ and $X_M{}^0/K^2$ plotted against frequency, at 25°C. The dashed line represents $\omega M/K^2$.

To insure contact between the samples and the plate and cylinders, the disks were usually compressed about ten percent after introduction, and allowed to stand several hours to permit the compressive stress to relax. Tests showed that the dynamic properties did not change during this relaxation, however. The polyisobutylene was sufficiently tacky to make excellent mechanical contact following this procedure; in fact, it was not easy to remove the disks after completion of measurements.

Table I gives an example of the magnitudes of R_M and $R_M{}^0$ and of η' calculated from their difference, as well as the corresponding quantities X_M, $X_M{}^0$, and G'. The magnitude of $R_M{}^0$ is never large enough so that its subtraction seriously affects the accuracy of η'. At low frequencies, $X_M{}^0$ is not a serious correction, but with increasing frequency its contribution to X_M of course becomes greater until their difference becomes too small to determine G' with adequate precision.

Values of G' and η' for the high molecular weight sample at three temperatures are plotted against the logarithm of the frequency in Figs. 5 and 6, and the same quantities for the low molecular weight sample in Figs. 7 and 8. In each case, two different sets of disks have been employed, with A/h ratios differing by a factor of two to four. The agreement of the values of G' and η' obtained, shown by the tagged points, justifies the use of Eqs. (4) and (6) within the frequency range employed.

Individual readings on the same sample are generally reproducible within about one percent, giving an error of about two percent in the values of R_M and X_M. Values for different samples show slightly more deviation, probably reflecting an additional error in the

the advantage that the compliances of coils and connector are included with K^2 in a single calibration which is of course a function of frequency. No assumption of additivity of impedances is necessary. The method requires addition of masses for calibration at the center of the connector; this is not possible with our present apparatus, whose construction was completed before the appearance of Trent's work.

[18] We are indebted to Dr. J. N. Ashworth and Mr. M. F. Johnson for the molecular weight values.

measurement of h, the sample thickness. Very few values for the pairs of samples shown in Figs. 5–8 are more than five percent apart, and this seems to be a safe figure to take for the reproducibility of the values reported here.

These data were obtained with a constant driving current of 10 ma; the peak amplitude and velocity of course varied depending on the total mechanical impedance and the frequency. The shear and rate of shear were always so small, however, that linear viscoelastic behavior[13,14] is to be expected, and the quantities measured should be independent of amplitude. In several cases this was confirmed by varying I_1 over a factor of four, with no resultant change in R_M or X_M.

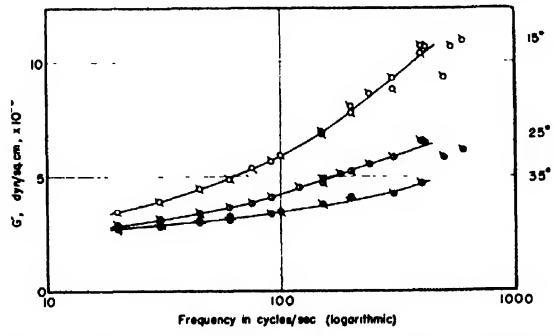


FIG. 5. Real part of dynamic rigidity of polyisobutylene, $\bar{M}_\eta = 1.2\times10^6$, plotted against the logarithm of the frequency. Upward tags, $2A/h = 20.3$; downward tags, $2A/h = 5.20$.

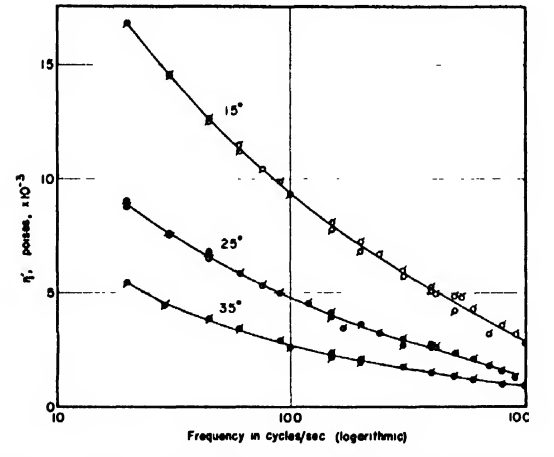


FIG. 6. Real part of dynamic viscosity of polyisobutylene, $\bar{M}_\eta = 1.2\times10^6$, plotted against the logarithm of the frequency. Key to tags same as in Fig. 5.

At each temperature and molecular weight, G' increases and η' decreases with increasing frequency. At any given frequency, both G' and η' decrease markedly with increasing temperature. An increase in molecular weight from 0.47 to 1.2 million increases G' by less than 20 percent, and has very little effect on η'.

DISCUSSION

There appear to be no other data in the literature on uncrosslinked rubbers with which the above results can be compared. It is of interest, however, to compare them qualitatively with similar data for vulcanized rubbers. Both G' and η' for polyisobutylene are found to be of the same order of magnitude as values for various cured rubbers (natural, GR-S, and butyl) at comparable frequency and temperature, derived from the measurements of Dillon, Prettyman, and Hall,[19] Nolle,[20] and Guth.[21] A quantitative comparison is difficult to make at present because chemical constitution has a considerable influence on the detailed behavior, though not apparently the order of magnitude, of G' and η'.

Dependence on Temperature

The variation of both G' and η' of polyisobutylene with temperature can be quantitatively explained by a single assumption, that stresses in the material relax by a variety of flow processes but the temperature de-

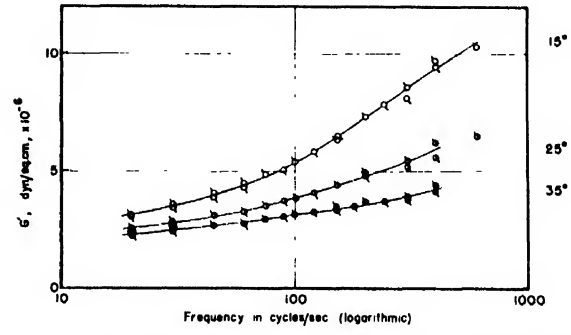


FIG. 7. Real part of dynamic rigidity of polyisobutylene, $\bar{M}_\eta = 0.47\times10^6$, plotted against the logarithm of the frequency. Upward tags, $2A/h = 24.5$; downward tags, $2A/h = 13.8$.

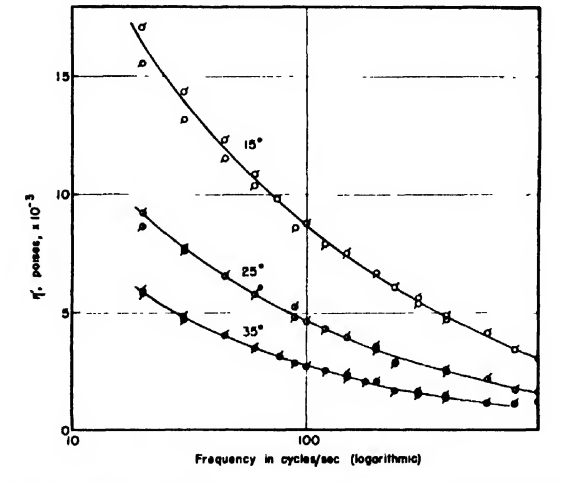


FIG. 8. Real part of dynamic viscosity of polyisobutylene, $\bar{M}_\eta = 0.47\times10^6$, plotted against the logarithm of the frequency. Keys to tags same as in Fig. 7.

[19] Dillon, Prettyman, and Hall, J. App. Phys. **15**, 309 (1944).
[20] A. W. Nolle, J. Polymer Sci. (to be published, and private communication); R. B. Blizard, Quarterly Progress Report of the Acoustics Laboratory, Massachusetts Institute of Technology (April-June, 1949).
[21] Ivey, Mrowca, and Guth, J. App. Phys. **20**, 486 (1949).

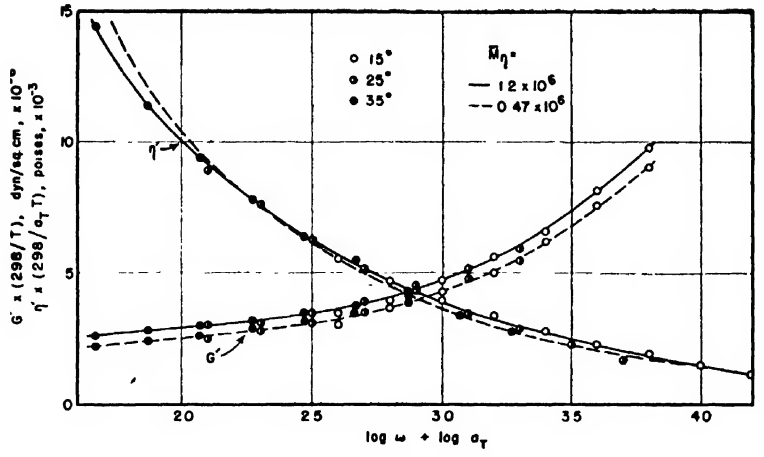


FIG. 9. Superposition of data at different temperatures: $G'(298/T)$ and $\eta'(298/a_T T)$ plotted against $\log\omega+\log a_T$. At 15°C, $a_T=3.16$; at 35°C, $a_T=0.38$. The points plotted for η' correspond to the solid curve ($\bar{M}_\eta=1.2\times10^6$); the points corresponding to the dashed curve ($\bar{M}_\eta=0.47\times10^6$) have been omitted to avoid confusion.

pendence of every such process is the same and identical with that of the steady flow viscosity. It is convenient to discuss the mechanical behavior in terms of an arbitrary distribution function of relaxation times, $G(\tau)$, such that the instantaneous rigidity is $\int_0^\infty G(\tau)d\tau$ and the steady flow viscosity is $\int_0^\infty \tau G(\tau)d\tau$. This has been employed by Andrews[3] and Kuhn,[22] and it is equivalent to the generalized Maxwell model of Alfrey and Doty;[23] that is, every spring $G(\tau)d\tau$ is relaxed by a dashpot $\tau G(\tau)d\tau$. In this case

$$G'=\int_0^\infty [G(\tau)\omega^2\tau^2/(1+\omega^2\tau^2)]d\tau$$

$$\text{and}\quad \eta'=\int_0^\infty [\tau G(\tau)/(1+\omega^2\tau^2)]d\tau.$$

Now, a change in temperature has two separate effects; it shifts the dispersion function to a different frequency region by altering the dimensionless variable $\omega\tau$, and it also changes the magnitudes of G' and η' by altering $G(\tau)$ and $\tau G(\tau)$, respectively. We will suppose that the rigidity mechanisms are of the nature of rubberlike elasticity and therefore that every $G(\tau)$ is proportional to the absolute temperature; this gives a slightly better fit with the data than does the assumption that the rigidity is independent of temperature, although the experimental temperature range is too small to test it critically. Then if a temperature change from T_0 to T (on the absolute scale) changes every relaxation time by a factor a_T, every contribution to G' will be multiplied by T/T_0 and shifted along a logarithmic frequency scale a distance $-\log a_T$; every contribution to η' will be multiplied by $a_T T/T_0$ and also shifted $-\log a_T$ along the $\log\omega$-axis. Moreover, the steady flow viscosity, η, will also be changed by the factor $a_T T/T_0$. Accordingly, choosing $T_0=298$°K, if we plot $G'(298/T)$ and $\eta'(298/a_T T)$ against $\log\omega+\log a_T$, data at all different temperatures should superpose.

Very successful superposition of both G' and η' for both molecular weights is obtained by selecting a_T values of 3.16 at 15°C and 0.37 at 35°C (Fig. 9). These values also fit in quite well with the temperature dependence of the steady flow viscosity of polyisobutylene given by Fox and Flory,[21] who showed that the apparent activation energy for viscous flow is independent of molecular weight and equal to 19.3 kcal. at 0° and 13.7 kcal. at 50°C. From the relation $\eta/\eta_{298}=a_T T/298$ we calculate 19 kcal. at 20° and 16.5 kcal. at 30°C. The values obtained by Andrews[3] from stress relaxation measurements are also from 15 to 16 kcal. in the range from 30° to 100°C. It appears therefore that all the flow processes in the relaxation spectrum which are reflected in these various measurements have the same temperature dependence.

Dependence on Molecular Weight

Figure 9 also emphasizes the very slight dependence of G' and η' on molecular weight. This indicates that the dynamic mechanical behavior involves primarily entanglements of molecular segments which are short compared with the total length, and that the loose ends of the molecules have relatively little significance. On this basis, it would be expected that η' would show still less dependence than G', as is indeed observed, because the contributions to η' come largely from mechanisms with short relaxation times (corresponding probably to shorter segments), while the contributions to G' come largely from mechanisms with long relaxation times (corresponding to long segments which could be perceptibly influenced by loose ends). By contrast, the steady flow viscosities of these two samples (if the viscosity average molecular weight can be assumed approximately proportional to the weight average) should differ by over a factor of ten,[24] and it is obvious from casual inspection of the samples that they do. This of course reflects the fact that the difficulty of moving entire molecules past each other increases markedly with molecular weight. Since η' must ap-

[22] Kuhn, Künzle, and Preissmann, Helv. Chim. Acta **30**, 307, 464 (1947).
[23] T. Alfrey and P. Doty, J. App. Phys. **16**, 700 (1945).
[24] T. G. Fox and P. J. Flory, J. Am. Chem. Soc. **70**, 2384 (1948).

proach η as $\omega \rightarrow 0$, the curves for η' must diverge widely at lower frequencies. Since η is of the order of 10^9 to 10^{10} poises,[24] and the maximum value of η' found here is about 10^4, the dispersion must extend well below the frequency range covered experimentally.

Dependence on Frequency

Figure 9 provides a picture of the dispersion functions for G' and η' over about two decades of frequency. There is at present no theoretical basis for interpretation of this dispersion. The relaxation distribution function derived by Kirkwood,[25] and the semi-empirical function of Kuhn,[22] are devised for crosslinked rubbers, and are not applicable to a linear polymer; tests have shown that neither fits our data for polyisobutylene. The "step" function used by Andrews[3] to describe stress relaxation in polyisobutylene provides (for frequencies not too close to the reciprocals of the cut-off times) that G' should be a linear function of $\log\omega$ and η' should be inversely proportional to ω; neither of these relations is even approximately fulfilled.

If a relaxation distribution function, either theoretical or empirical, can be obtained to fit our dynamic data, it should also apply to the transient stress relaxation data of Andrews,[3] with the reservation that his experiments involved extension rather than shear and represented a far greater extent of strain. However, a considerable gap in time scale remains to be bridged between the shortest transient measurement (about ten seconds) and the longest period in dynamic measurement (about one-twentieth of a second). Experimental data in this range are needed to provide a complete picture from which the mechanism of elasticity and elastic loss in uncrosslinked rubbers may be deduced.

[25] J. G. Kirkwood, J. Chem. Phys. 14, 51 (1946)

ACKNOWLEDGMENT

This work was supported in part by a grant from Research Corporation, and in part by the Research Committee of the Graduate School of the University of Wisconsin from funds supplied by the Wisconsin Alumni Research Foundation. We are deeply indebted to Mr. Frederic W. Schremp for help in the detailed design of the apparatus, and to Mr. Thomas R. Lloyd for painstaking work in its construction. Professor F. T. Adler and Professor W. T. Thomson have given valuable advice in problems relating to mechanical resonances in the apparatus. We are grateful to Carbide and Carbon Chemicals Corporation for support through its Fellowship in Physical Chemistry.

The Dielectric Strength of Gaseous Fluorocarbons

W. A. Wilson, J. H. Simons, and T. J. Brice

Fluorine Laboratories, The Pennsylvania State College, State College, Pennsylvania

(Received August 16, 1949)

The sixty-cycle dielectric strengths of some gaseous fluorocarbons have been measured between three differently shaped electrode pairs at pressures up to three atmospheres. The breakdown potentials for propforane, butforane, and pentforane were found, in most instances, to be equal to or greater than those for sulfur hexafluoride under comparable conditions, and to be far greater than those for nitrogen. Fluorocarbons thus have possible uses as gaseous insulators in high voltage apparatus.

A CONSIDERABLE amount of research has been carried out during the past few years in evaluating the dielectric strengths of various gaseous materials. Rodine and Herb[1] found that if carbon tetrachloride vapors were added to air or nitrogen, the dielectric strength was raised considerably. Charlton and Cooper[2] measured the dielectric strength of numerous gases and found that the Freon compounds had much higher dielectric strengths than nitrogen under the same experimental conditions. Camilli and Chapman[3] have found that sulfur hexafluoride and dichloromethforane have dielectric strengths much higher than nitrogen, and, when the gases were under about three atmospheres pressure, the dielectric strengths of these compounds approached values obtained for transformer oils. These results have been utilized commercially by insulating certain high voltage condensers, transformers, and electrostatic generators with various compressed gases.

During this research, the sixty-cycle dielectric strengths of various fluorocarbon gases and vapors were measured. Propforane, butforane, and pentforane were found, in most instances, to have dielectric strengths equal to or greater than sulfur hexafluoride under similar conditions.

The experiments were carried out in three glass cells containing copper electrode pairs of different shapes. One cell contained flat, circular disks of 0.477-inch diameters, the second had hemispheres of 0.5-inch radii, and the third had a hemisphere of 0.5-inch radius and a pointed 0.149-inch rod as electrodes. The pointed

[1] M. D. Rodine and R. G. Herb, Phys. Rev. 51, 508 (1937).
[2] E. E. Charlton and F. S. Cooper, Gen. Elec. Rev. 40, 438 (1937).
[3] G. Camilli and J. J. Chapman, A.I.E.E. Technical Paper 47-240 (September, 1947).

end was a section of a right cone with 15°27′ semiangle and bases having diameters of 0.036 inch and 0.149 inch. The electrodes were sealed symmetrically about common axes into round bottom Pyrex flasks and subsequently sealed to a portable vacuum system. A 100 ml flask was used for the disk electrodes, and one liter flasks were used for the other cells. These electrode shapes are similar to those generally used for this type of measurement.[3] The pair of hemispheres gives an approximately uniform field, while the hemisphere-rod arrangement gives a very non-uniform field.

A small laboratory transformer was used for producing voltages less than 35 kv. For higher voltages, a transformer capable of producing 150 kv was used. The voltage measurements were made with an accuracy of ±5 percent.

The fluorocarbons were prepared by the electrochemical process described by Simons.[4] The samples used were fractions of narrow boiling range from distillations through efficient columns. Molecular weight values were within ±0.5 percent of the theoretical values. The samples were dried over P_2O_5 before distillation and again just prior to the dielectric strength measurements.

At the dielectric breakdown potential, arcing between the electrodes occurs. It was noticed that, except in the case of methforane, this arcing in fluorocarbons resulted in the formation of a thin layer of black material, probably carbon, on the surfaces of the electrodes and the glass flask; and that a pressure rise of about one per cent of the total pressure occurred. This latter observation seemed to indicate that lower molecular weight fluorocarbons formed to a slight extent during arcing. Since the last of a series of five breakdown potentials on a given gas sample was essentially the same as the first, it was evident that small amounts of these materials did not noticeably affect the dielectric strength of the fluorocarbon. In addition, the dielectric breakdown potentials of fresh gas samples were found to be the same, within experimental error, in a cell coated with the solid as in a new, uncoated cell.

The dielectric strengths are plotted against pressure on Figs. 1 to 5. Each plot represents the results obtained using a different electrode pair or a different electrode gap spacing. Sulfur hexafluoride and nitrogen were tested for comparison.

The tests using the disk electrodes (Fig. 1) were carried out at a gap spacing of 0.125 inch. The dielectric strength of the fluorocarbons increased with increasing molecular weight, and all the fluorocarbons had values higher than those for nitrogen. Butforane and pentforane had dielectric strengths higher than sulfur hexa-

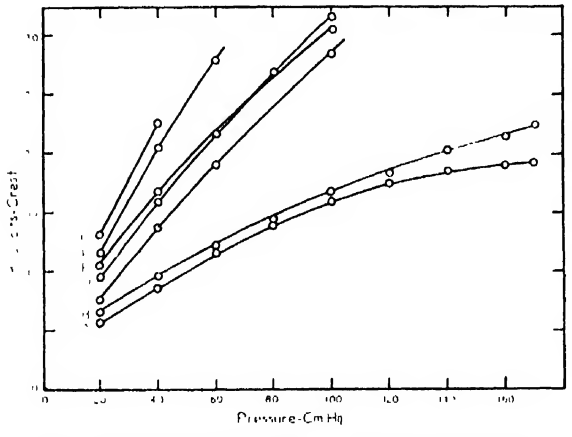


FIG. 1. 60-cycle dielectric strength between 0.477-inch diameter disks: Gap; 0.125 inch.

A—Nitrogen
B—Methforane
C—Ethforane
D—Propforane
E—Sulfur Hexafluoride
F—Butforane
G—Pentforane

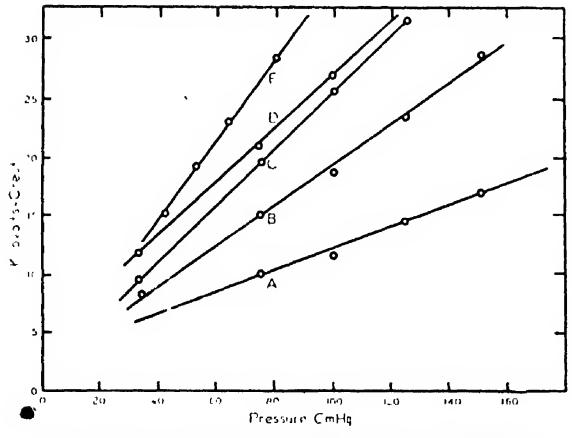


FIG. 2. 60-cycle dielectric strength between 0.500-inch radii hemispheres: Gap; 0.100 inch.

A—Nitrogen
B—Ethforane
C—Propforane
D—Sulfur Hexafluoride
E—Butforane

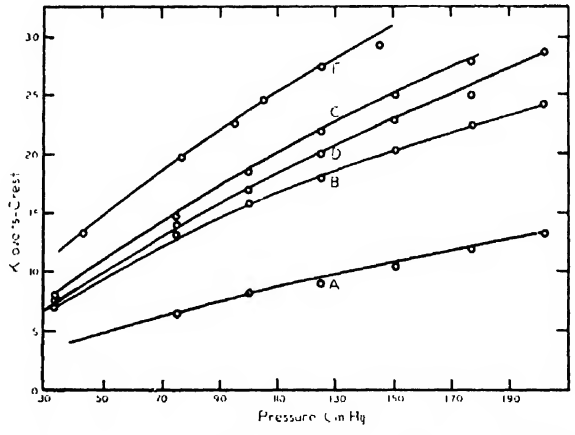


FIG. 3. 60-cycle dielectric strength between a 0.500-inch radius hemisphere and a pointed 0.149-inch diameter rod: Gap; 0.100 inch.

A—Nitrogen
B—Ethforane
C—Propforane
D—Sulfur Hexafluoride
E—Butforane

[4] J. H. Simons and co-workers, J. Electrochem. Soc. **95**, 47 (1949).

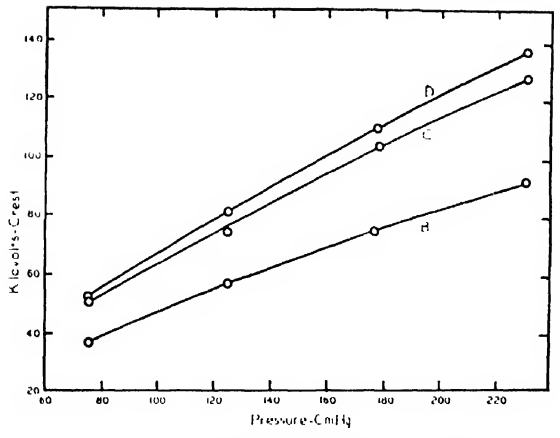


FIG. 4. 60-cycle dielectric strength between 0.500-inch radii hemispheres: Gap; 0.25 inch.

B - Ethforane C—Propforane
D—Sulfur Hexafluoride

fluoride. The dielectric strength of propforane was lower than that of sulfur hexafluoride at pressures less than one atmosphere and higher than that of sulfur hexafluoride at pressures above one atmosphere.

The tests using the pair of hemispheres (Figs. 2 and 4) were carried out at gap spacings of 0.100 inch and 0.250 inch. The breakdown potentials obtained at equal pressures were roughly proportional to the gap spacing. The dielectric strength of propforane was somewhat lower than that of sulfur hexafluoride at all pressures used. The general pattern of relative dielectric strength values for the other compounds was the same as obtained with the pair of disks.

The tests using the hemisphere-point electrode pair (Figs. 3 and 5) were also carried out at gap spacing of 0.100 inch and 0.250 inch. As was expected, the breakdown potentials obtained when using this electrode pair were lower than those obtained when using the symmetrical electrode pairs which produced more uniform fields. When tested with this pair of electrodes, the dielectric strength of propforane was higher than that of sulfur hexafluoride at all pressures used. The pattern of relative values for the other compounds was the same as was obtained with the other electrode pairs.

An interesting correlation of the dielectric strengths of fluorocarbons is illustrated in Fig. 6. The dielectric strengths of fluorocarbons (taken from Fig. 1) are plotted against their gas densities. As a first approximation, the dielectric strength is proportional to the gas density, independent of the compound. Similar treatment of the data obtained when using the other electrode pairs gave similar results.

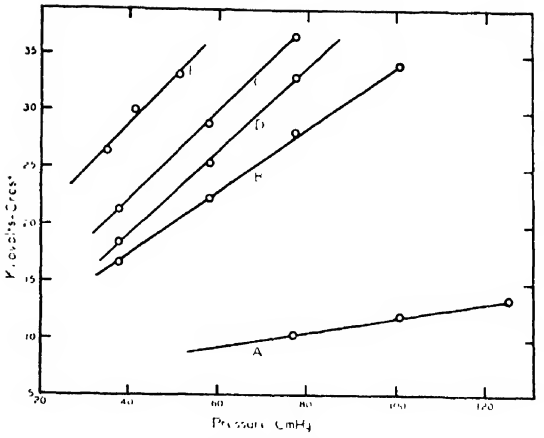


FIG. 5. 60-cycle dielectric strength between a 0.500-inch radius hemisphere and a pointed 0.149-inch diameter rod: Gap; 0.25 inch.

A—Nitrogen C—Propforane
B—Ethforane D—Sulfur Hexafluoride
E—Butforane

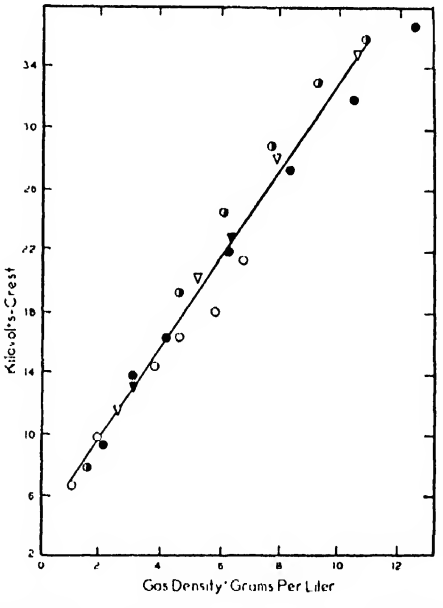


FIG. 6. 60-cycle dielectric strength between 0.477-inch diameter disks: Gap; 0.125 inch.

○ Methforane ● Propforane
◑ Ethforane △ Butforane
▲ Pentforane

ACKNOWLEDGMENT

The authors wish to thank Professor R. E. Armington of the Electrical Engineering Department of The Pennsylvania State College for his assistance in obtaining the measurements reported here and the Electrical Engineering Department for allowing the use of their high voltage transformer.

This work was done under the sponsorship of the Minnesota Mining & Manufacturing Company, St. Paul, Minnesota.

Complex Stressing of Polyethylene

I. L. Hopkins, W. O. Baker, and J. B. Howard
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received August 17, 1949)

Polyethylene of such molecular weight and structure that it readily fibers or cold draws to 300–600 percent elongation by usual uniaxial, tensile stressing may react quite differently under biaxial tension. When biaxial tension in 1:1 ratio is applied to a diaphragm, some polymers show brittle fracture with <20 percent elongation at break. However, if the average molecular weight of such polyethylenes is shifted upward by crude fractionation, or an initially higher average is used, the polymers orient under complex stresses. Then, they usually elongate several hundred percent before rupture. Variations in crystallinity are also significant, although most technical polyethylene soon attains at room temperature enough crystallinity so that this factor does not cause big differences.

Although the whole study is so far preliminary, x-ray scattering of stressed samples suggests that preferred glide on certain crystallite planes tends to occur as the yield point approaches. These are such as to inhibit smooth alignment of the long chain axis in the direction of stressing. This could lead to brittleness.

Apparatus for complex stressing of sheets and tubes is described. Strains are taken from coordinates printed on the sample by the silk screen process. High speed stressing was also observed. The speed of retraction of amorphous polyethylene chains approaches that of rubber.

THE usual tensile strength or stress-strain measurement for plastics and rubbers involves uniaxial stretching of a long, thin strip. With such tension, polyethylene, and other microcrystalline polymers such as the polyamides and polyesters, give a steep initial stress *vs.* strain curve, up to 20–50 percent elongation. Then, "yield" occurs, and a quite flat curve delineates cold-drawing or fibering, marked by crystallite orientation. Finally, between 400 and 700 percent elongation, depending on the polymer, rate of stressing, degree of crystallinity, etc., the curve bends upward again and rupture soon occurs.[1,2]

However, two-directional or biaxial stress-strain properties have also been observed for rubber[3,4,5,6] and for thin Nylon films.[7] In these cases, while the form of the stress-strain curve is different than with uniaxial tension, the strains at rupture have apparently the usual values—of several hundred percent.

With polyethylene, we have found that biaxial stressing of some commercial grades (determined by some function of average molecular weight and chain linearity, such as melt viscosity), causes brittle fracture at 30 percent elongation or less. This is with a polymer having simple tensile properties of curves *C*—uniaxial in Fig. 1, wherein it breaks only after 300–500 percent elongation. Somewhat higher molecular weight polymer, however, shows much higher elongation at rupture, perhaps even approaching the lower uniaxial values. An example of this appears in curves *A* of Fig. 1. In these curves, the end point does *not* indicate rupture, as it does in curves *C*, except for the uniaxial tension curves *A*. Rather, on curves *A* for biaxial tension, the end of the curve denotes yield, where an oriented bubble begins to form on the stressed diaphragm. Accordingly, the material of curves *A* begins to orient at ~70 percent elongation, and draws down to 200 percent or more in the bubble before rupture. Nevertheless, its uniaxial curve is seen, in Fig. 1, to come quite close to that of material *C*.

The relative differences in average molecular weight of these two polymers are represented in Table I, by their inherent viscosities.

$$(\ln\eta_r)/C$$

in *m*-xylene (90 percent) at 85°C, with $C=0.5$ g/100-cc solution, and with η_r = time of flow of solution/time of flow of solvent.

The inherent viscosity values agree with the *uniaxial* stress-strain data of Fig. 1 in showing that both materials *A* and *C* are well within what is commonly called the "super-polymer" "cold drawing" range. Biaxial stresses thus have brought out a wholly unexpected and dramatic sensitivity of this sort of me-

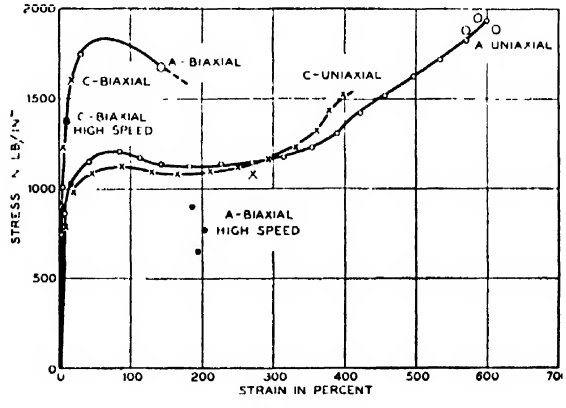


FIG. 1. Stress-strain curves of polyethylenes *A* and *C* under uniaxial and biaxial tension. (Large circles (for *A*) and crosses (for *C*) are rupture points, except that the large circle for *A*-biaxial is the yield point.)

[1] Hahn, Macht, and Fletcher, Ind. Eng. Chem. **37**, 526 (1945), show a fairly typical stress-strain curve for polyethylene.
[2] E. Hunter and W. G. Oakes, Brit. Plastics **17**, 94 (1945).
[3] P. Prache, Caoutchouc et gutta-percha **27**, 15, 125 (1930).
[4] J. R. Sheppard and W. J. Clapson, Ind. Eng. Chem. **24**, 782 (1932).
[5] C. F. Flint and W. J. S. Naunton, Trans. Inst. Rubber Ind. **12**, 367 (1937).
[6] L. R. G. Treloar, Trans. Faraday Soc. **40**, 59 (1944); Trans. Inst. Rubber Ind. **19**, 201 (1944).
[7] J. Miklowitz, J. Colloid Sci. **2**, 193, 217 (1947).

chanical behavior to average molecular weight as reflected in over-all molecular cohesion. Interestingly, the inherent viscosity for polymer *C* is actually higher than that for polymer *A*. The melt viscosities are reversed, however, with polymer *A* higher; it may be a more branched material although the over-all degreees of crystallinity of the two as judged by x-ray scattering, are similar. Further, polymer *A* has a definitely smaller fraction of low molecular weight, soluble in boiling chloroform, than polymer *C*; this is important, as judged from scores of similar comparisons made in this study.

For instance, polymer *D* in Table I showed, like polymer *C*, low elongation and brittle fracture under biaxial tension. It was then dissolved in pure trichloroethylene, and the system equilibrated at 43.8°C. Sixty-one percent by weight of the original sample was insoluble at this temperature; it had the properties of *D'* in Table I. The fraction of *D'* soluble in boiling chloroform was reduced fourfold from that of *D*, and the polymer was converted to high elongation cold-drawing behavior. However, its stress at yield was no higher than that of the unfractionated material (at rupture). By contrast, polymer *D''* represents the 25 percent of polymer *D* insoluble in trichloroethylene between 43.8° and 25°C. Its density is a little higher than that of *D'*; its inherent viscosity is a little higher than that of the parent *D*; this may mean more chain linearity and crystallinity. But the portion soluble in boiling $CHCl_3$ has risen to 19 percent and the ultimate elongation under biaxial stress has fallen to 28 percent, accompanied by brittle fracture.

A', in Table I, reflects still another interesting aspect of biaxial mechanics. It is the fraction, 91.5 percent of polymer *A*, insoluble in the benzene-isopropanol azeotrope at its boiling point. The portion of *A'* soluble in boiling $CHCl_3$ is tiny, the density is still low, however, as well as inherent viscosity—some balance between removal of easily soluble species, but retention of (presumably) branched species appears. Hence, the material exhibits the orientation and tenacity of the parent *A*, but the elongation at yield *drops* sharply. This is a very typical behavior of "tough" polyethylene, and should not be confused with an approach to the low elongation at simultaneous yield and rupture of the polymers showing brittle fractures. In other words, as a function of average molecular weight and structure, the elongation at yield of polyethylene goes through a maximum, while the ultimate tenacity increases continuously.

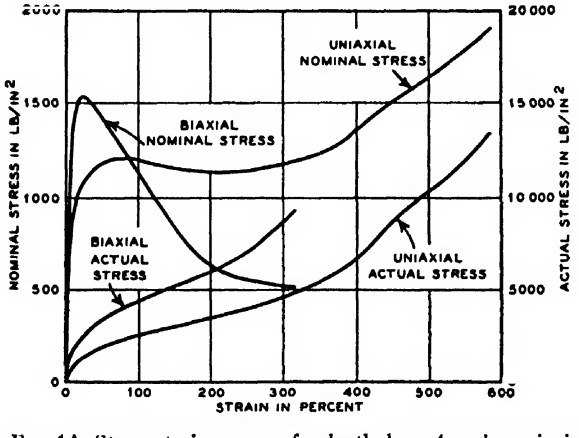


FIG. 1A. Stress-strain curves of polyethylene *A* under uniaxial tension, and biaxial tension with liquid medium. The biaxial nominal stress curve passes through and somewhat beyond the area of rupture points determined by the high speed tests with nitrogen as the pressure medium.

TABLE I. Properties of polyethylene under biaxial tension.

Polymer	Density 25°C	$\ln\eta_r/C$	Percent soluble in boiling chloroform (N.F.P.)*	Strain at "yield" %	Strain at rupture %	Stress at "yield" p.s.i.
A	0.9191	0.933	9.6	70	~200	1600
A' (=91.5% of *A*)	0.9171	0.974	1.9	44	~200	1670
C	0.9197	1.007	14.5	17	17	1600
D	0.9167	0.879	12.6	24	24	1480
D' (=61% of *D*)	0.9191	1.007	3.8	66	~200	1450
D'' (=25% of *D*)	0.9197	0.902	19.0	28	28	1360

* Amount soluble under empirical conditions.

This would have to be established finally by detailed fractionation of polyethylene; this has been explored, but not solved for reasons well recognized.[8] However, some additional examples of reaction of unfractionated, commercial-type polyethylenes to biaxial stress appear in Fig. 2. Apparently, with these molecular weight distributions, brittle fractures and low elongations do not occur much above an inherent viscosity of 1.0. Then, the above effect, where the still higher molecular weight polymers again show reduced elongations at yield, although they orient highly before rupture, appears at the extreme right in Fig. 2. Many additional points to those in Fig. 2 have been obtained, but they still do not allow construction of what might be a convex-upward correlation curve. A new correlation technique,[9] which has been especially useful in experimental studies of this kind, shows that extreme, or peripheral, association is probable in Fig. 2, however.

EXPERIMENTAL METHOD

Biaxial stresses were applied by N_2 gas pressure on a polymer disk. The cell in which the disks, in the range 0.080-in. to 0.140-in. thick, were clamped is shown in Fig. 3. The sample was illuminated by floodlights, behind heat-absorbing glass plates. It was photographed during straining, sometimes with a movie camera, but usually at intervals with the fixed camera shown behind the lights. Stress measurements were recorded by holding up a card bearing a definite figure for gas pressure (250 is shown in figure) and photographing the

[8] R. B. Richards, Trans. Faraday Soc. 42, 10 (1946).
[9] P. S. Olmstead and J. W. Tukey, Ann. Math. Statistics 18, 495 (1947).

card and the strained specimen just when the gas gauge reached the given figure.

The specimens were either clear polyethylene or a compound containing 1 percent or 2 percent by weight of fine carbon black, which had no effect on these physical properties. They were compression molded so as to be isotropic and free from included stresses. At first, 0.10-in. coordinate grids were photographed on them after coating with emulsion. However, it was found that the silk screen method of printing, long used for posters and textiles, worked very well with either a black or white ink containing a chlorinated rubber base. The ink had no effect of either crazing or swelling the

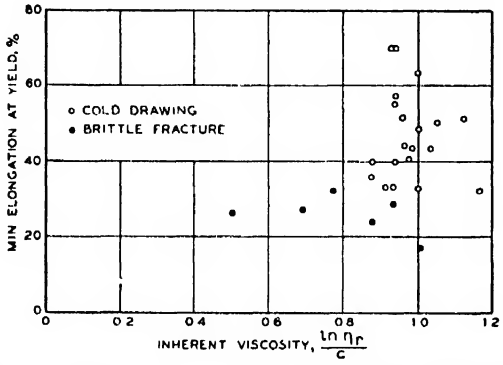


FIG. 2. Minimum elongation at yield under biaxial tension plotted against inherent viscosity of dilute solutions of polyethylene of various average molecular weights.

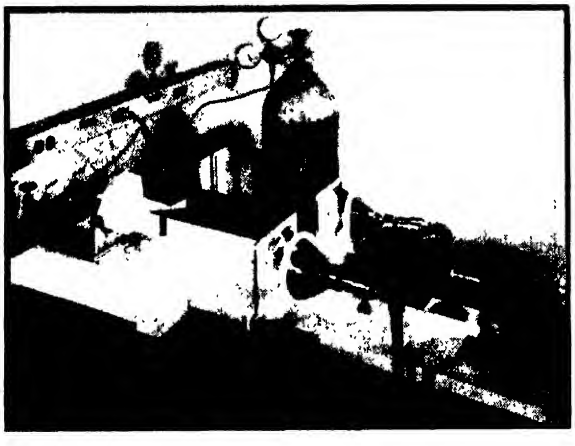


FIG. 3. Experimental arrangement for determining biaxial stress-strain curve. Diaphragm of polymer, bearing coordinates, appears below sign showing that it was subject at instant it was photographed (camera behind floodlights) to gauge pressure of 250 p.s.i.

polymer. These coordinates appear on the specimen shown in Fig. 3. Figure 4 shows typical drawn (upper and lower left) *vs.* brittle (lower right, with stick to mark crack) fractures. The white disk in Fig. 4 showed the reduced elongation at yield encountered for samples of high melt viscosity. Figure 3 shows a mirror obliquely placed along side the specimen holder so that the bulge of the diaphragm could be measured on the photograph along with the expansion of coordinates. From this, the stress in p.s.i. was computed as

$$PR/2t,$$

P being gas pressure in p.s.i., R, the radius of the (spherical) bulge and t, the thickness of the sheet in inches. Strain was obtained directly from the expansion of coordinates, which was the same in both directions (1:1 ratio) in most experiments. Some measurements were also done in a device in which cylindrical tubes of the polymer were blown up. This apparatus is shown in Fig. 5, with a polyethylene tube of 1 in. outside diameter and $\frac{1}{16}$-in. wall thickness. Coordinates were again printed directly on the specimen. Here, the principal stresses have a 2:1 ratio, but the results were quite comparable to the 1:1 ratio situation.

The rate of stressing was such that the pressure gauge rose uniformly over a period of 60 ± 10 sec. followed by abrupt rupture. Some experiments ten times as slow gave similar results and stress relaxation became significant for long periods; the present results suggest a study of stress relaxation under biaxial tension compared to the usual uniaxial case. Also, effects from very rapid straining were investigated. It was particularly desired here to be able to photograph the strains on the bubble which appeared as the high elongation samples oriented. This bubble is what blew into the crater shown on the upper sample of Fig. 4. Even with slow rates of over-all stressing with gas pressure, this bubble always appeared and burst so fast that ordinary photographs did not slow its growth. Hence, high speed

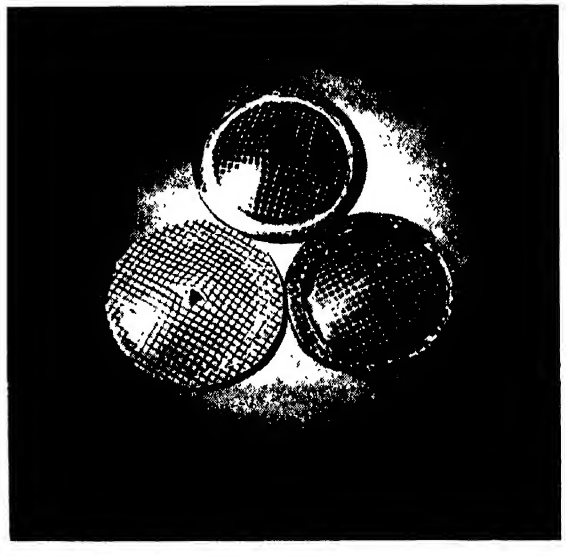


FIG. 4. Polyethylene disks ruptured by biaxial stress. Upper, moderate molecular weight, high yield strain, extensive cold drawing. Lower left, high molecular weight, lower yield strain, extensive cold drawing. Lower right, marginal molecular weight, low "yield" strain coinciding with brittle rupture, no cold drawing (stick is in crack to force it open for view).

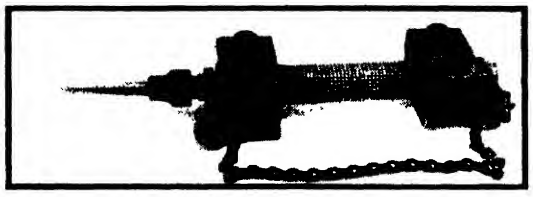


FIG. 5. Device for applying gas or hydraulic pressure to polymer tube bearing coordinates. Tube is clamped into fittings within steel blocks. (Here, rupture often occurs in position away from a single camera.)

photographs, with the Fastax camera, were made of specimens stressed and ruptured by sudden opening of the gas valve. The whole time from initial pressure to breaking was about 0.15 sec.; the bubble formed so rapidly that its life was roughly 10^{-3} sec. However, a few measurements of the coordinates on the bubble could be made, both as it formed and as the diaphragm relaxed after bursting. They are shown on Fig. 1 in terms of stress-strain assuming (1) that the stress-strain curve up to the point of yield (bubble formation) or brittle fracture is the same as in slower tests, (2) hence the pressures are the same at this point and in the case of the formation of a bubble do not change appreciably over the brief time interval required for the formation and bursting of the bubble, (3) that from the measured radius of the bubble and the assumed pressure, the stress in the bubble may be computed and (4) the strain in the bubble (ultimate elongation of the material) can be approximated by the over-all bubble dimensions. More appear on Fig. 6, where strains directly measured from the Fastax frames are plotted against time. The recovery from 130 percent back to about 50 percent total strain takes about 6 milliseconds. This is the same order·velocity as for free contraction of soft rubber.[10,11] Probably amorphous chains of polyethylene are responsible; they are acting kinetic theory-wise compared to the usual sluggish "plastic" behavior of the polymer. In this connection, it will be noted later that biaxially stretched polyethylene seems to show an unusual orientation of the amorphous components (compared to their only slight orientation in an ordinary drawn fiber).

The high speed data still show the large relative differences in the polymers *A* and *C*. *C* breaks at 20 percent elongation, with brittle fracture, while *A* yields at a higher value, and cold draws to about 150–200 percent, as the points on the graph show, before breaking. Accordingly, failure of polymer *A* was not converted to brittle fracture by an "impact rupture" 10^4 times faster than the usual test. Incidentally, the growth of the brittle crack in polymer *C* could just about be seen in 10^{-4} sec. None of these data means, of course, that the apparent tenacity of the polymer is insensitive to the total *time* of stressing[12,13] (as in fatigue).

A modified apparatus is now being studied in which water (or aqueous anti-freeze solutions for low temperatures) is being used for pressure instead of gas. In this way, very little energy is stored during stressing, and the bubble formation for samples which orient, as well as the bulging of all samples, can be better controlled. Experiments done with water instead of nitrogen as the pressure transmitting medium bring test specimens

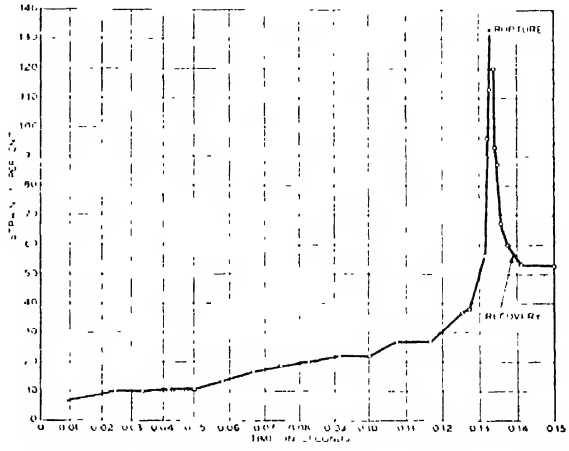


FIG. 6. Rate of straining of polyethylene diaphragm suddenly blown out by high gas pressure. The blown bubble itself remained highly oriented; recovery is for the rest of the surface.

gradually to the rupture point, rather than merely to the yield point, and show that biaxial strains of over 400 percent may be attained. Stress-strain curves for polyethylene *A* are shown in Fig. 1A. The biaxial nominal stress-strain curve of Fig. 1A passes through the biaxial high-speed rupture points of Fig. 1, confirming qualitatively the assumptions under which these points were calculated.

Reason for Biaxial Brittleness

Pending work in progress, and examination of other materials, only tentative qualitative explanations can be given for dependence of the elongation of polyethylene on stress complexity. In contrast to an assembly of long chain molecules like amorphous rubber, polyethylene at 25°C contains a major volume of small crystallites (100–300A in dimensions) and a minor volume of disordered matter. Hence, response of these crystallites to external stress, rather than gradual progressive orientation of individual chains as when rubber is stretched, is the critical process which will determine whether brittle fracture or cold drawing will occur. Of course, ultimately the long chain axes themselves line up in the direction of uniaxial fibering, as the

[10] R. B. Stambaugh, Phys. Rev. **65**, 250 (1944).
[11] Mrowca, Dart, and Guth, Phys. Rev. **66**, 32 (1944).
[12] R. N. Haward, Trans. Faraday Soc. **38**, 394 (1942); **39**, 267 (1943).
[13] T. Alfrey, *Mechanical Behavior of High Polymers* (Interscience Publishers, Inc., New York 1948), p. 481 ff.

crystallites are so oriented.[14] The mechanism of this adjustment,[15] the "shoulder" on cold drawn strips,[16] the dimensional changes on drawing, and stress-strain variations on redrawing[16] have also been considered in terms of crystallinity. But, less attention has been given to the striking structural changes in the early part of the uniaxial stress-strain curve, where it begins to flatten out on Fig. 1. Here, the long chain axes in the crystallites are actually put *at a large angle* ($\sim 70°$) *to the direction of external stress*[17,18] by the preferred glide on certain crystal planes. Thus, in the early stages of orientation, the x-ray reflections from 110 planes are split, up off the equator, while the other strong feature, the 200 reflections concentrate down on the equator.[17] At higher temperatures, where thermal agitation has disordered the crystallites[15,17] this anomalous orientation is absent,[18] and the long chain axes progressively align with the direction of stress.

Now, if such anisotropy of slip as this occurs during uniaxial pulling, it seems that under biaxial tension, a skewed orientation would result in which much of the pull between the chains was lateral. This would tend to pull them apart in the easiest way (since van der Waals bonds would support most of the stress). Hence, dependence of strength on molecular weight would be even more critical than in the uniaxial case.

Of course the "anomalous" orientation found at the beginning of uniaxial stressing of ordinary specimens (1 mm or more thick) may already reflect complex stressing. This is because of the appreciable shear attending stretching of such "thick" samples. An electron diffraction investigation of stretched polyethylene film initially about 1000A thick showed only progressive, normal, orientation of 110 planes,[19] but these films had some orientation during preparation. It would be interesting to see if such thin films, initially strictly isotropic, showed "anomalous" orientation from just stretching at 25°C.

In any case, x-ray scattering of polyethylene which has been sheared, or biaxially stretched (with some attendant shear), should show this competitive or "blocked" crystallite gliding. It does appear in the diagrams of Fig. 7. First, Fig. 7A shows the usual fiber diagram of highly, uniaxially, oriented polyethylene. The strong 110 and 200 reflections are well concentrated on the equator. The pattern (all are with crystal-monochromatized Cu $K\alpha$-radiation) also clearly exhibits the *unoriented*, amorphous components of the polymer.

Figure 7B is the diagram from a small sample roughly normal to the plane of principal shear in a tube which was twisted to failure. This tube was

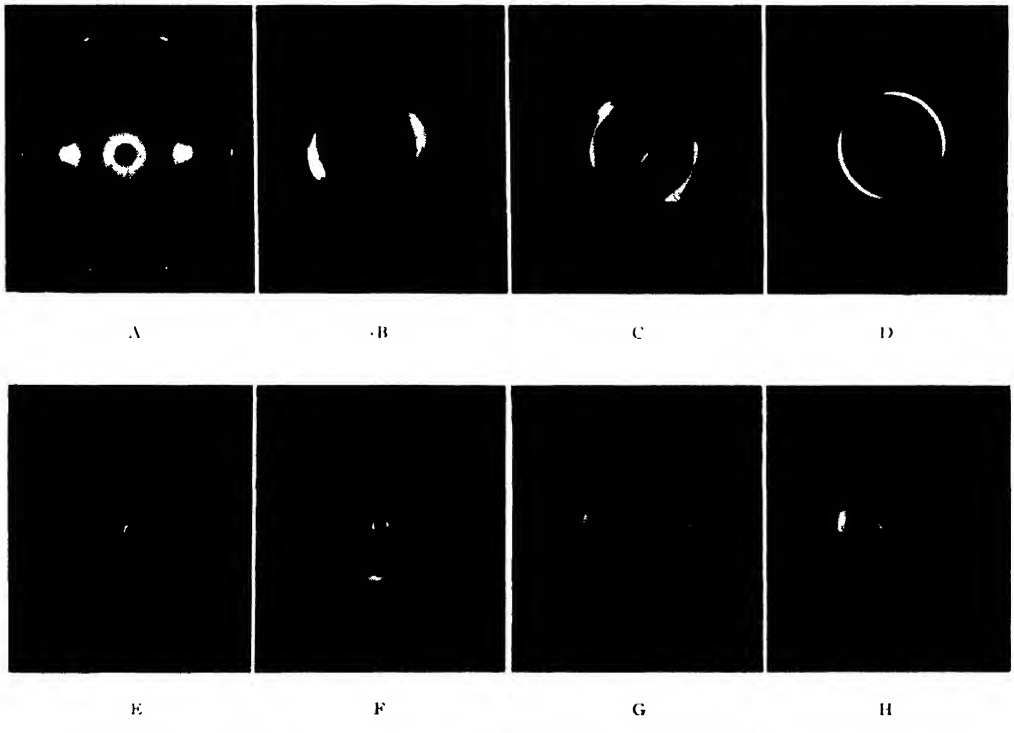


FIG. 7. X-ray scattering of polyethylene oriented by various kinds of stresses, as discussed in text.

[14] C. W. Bunn, Trans. Faraday Soc. **35**, 482 (1939).
[15] C. W. Bunn and T. C. Alcock, Trans. Faraday Soc. **41**, 317 (1945).
[16] W. M. D. Bryant, J. Polymer Sci. **2**, 547 (1947).
[17] W. O. Baker, *High Polymers* edited by Twiss (Reinhold Publishing Corporation, New York, 1945) p. 108.
[18] A. Brown, J. App. Phys. **20**, 552 (1949).
[19] A. Charlesby, Proc. Phys. Soc. **57**, 496 (1945).

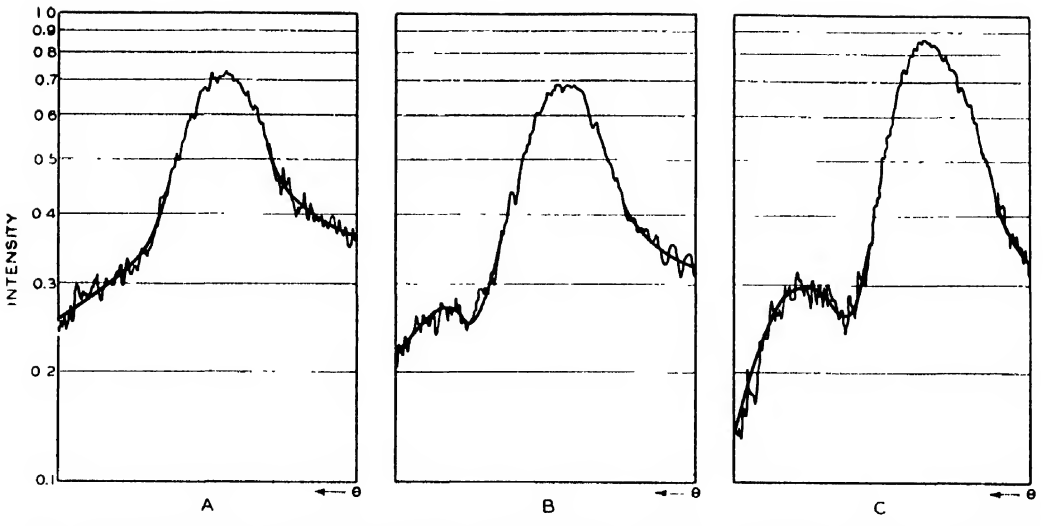


FIG. 8. Microphotometer tracings (uncorrected) showing x-ray scattering of principal features of unoriented polyethylene (A) immediately after quenching from melt to −78°C., (B) after same sample has stood 24 hr. at 25°C, (C) after annealing below melting point, and coming slowly at 25°C.

½ in. outside diameter, ¼ in. inside diameter, and about 12 in. long. It was twisted from the ends and sheared through near the middle. The B sample was at the rupture, and shows roughly normal alignment, although with a wide angular distribution and tapering of 110 and 200 intensities. Figure C is quite different, however. It shows what the polymer went through in getting to the final orientation just before rupture. It was obtained from a sample just back of the B sample. The 110 reflections are split, and also skewed with respect to the 200. The long chain axes by no means smoothly and gradually oriented in the direction of the principal shearing stress. However, Fig. 7D, whose sample came from a region of still lower shear than the C sample, shows that from the very outset these planes started to orient without "blocking."

Figures 7E–H exhibit these effects for specimens adjacent to the rupture of biaxially stressed disks which cold-drew or "bubbled," as shown at the top of Fig. 4. Samples were cut from the lips of the large circular bubble like that in Fig. 4. They were cut as little strips whose length lay along a radius of the test disk. Accordingly, they were somewhat tapered in thickness because of the drawing down at the rupture. When mounted with long axis vertical (like the fiber diagram of Fig. 7A), and the plane of the strip (and thus disk) normal to the x-ray beam, patterns like Fig. 7E result. From this point of view, the chain axes lie as though they had tried to form rings around the pole of the inflated disk. Figure 7F shows a similar diagram for a sample from a higher molecular weight polymer, such as in the white disk in Fig. 4. Here the stress at rupture was about 2000 p.s.i., as compared to 1500–1600 p.s.i. for the black, drawn disk. The orientation was more complete, but otherwise similar. When these radial strips were turned 90° around their length, so that the x-ray beam entered parallel to the plane of the strip (and the disk), diagrams like Fig. 7G were found. These indicate the "two-way" or cross-hatched orientation which has occurred in the plane. This is presumably because the crystallites have attempted to glide in the directions of principal stresses, but have been blocked because of the gliding anisotropy noted before. This results, again, in much of the external tensions being *across* chain axis and unless the molecular weight is quite high, remarkable brittle, unoriented fractures such as those of the third disk in Fig. 4 occur. The off-equator spacings in Fig. 7G are the familiar 4.17A and 3.70A of the *n*-paraffin lattice. The equatorial spacings are, however, the strong 4.17A (110) and a fuzzy fea-

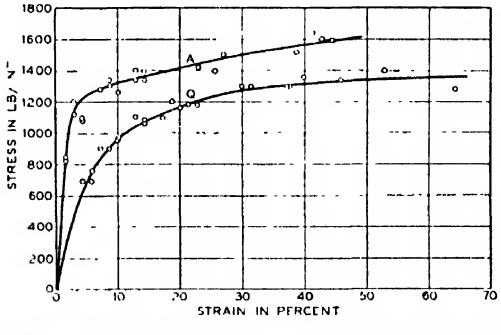


FIG. 9. Stress-strain curves of annealed and quenched polyethylene.

ture which has a Bragg spacing of 4.6A—presumably *oriented amorphous* chains whose unoriented halo scattering appears in Fig. 7A. This stretching out of amorphous chains may account for the quick, rubber-like contraction noted in connection with Fig. 6. Figure 7H shows the orientation appearing when the x-ray beam is parallel to the plane of and directed along the radius of the disk at the rupture.

DEGREE OF CRYSTALLINITY AND COMPLEX STRESSING

If preferred crystallite glide is so significant in reaction of polyethylene to complex stressing, large variations in crystallinity should be reflected in stress-strain properties. Degree of crystallinity of essentially simple chain polymers may be influenced in two ways. First, equilibrium crystallinity may diminish (or transform) with increase in temperature.[15,17,20,21] Secondly, *metastable* reduction in crystallinity occurs in quenched or

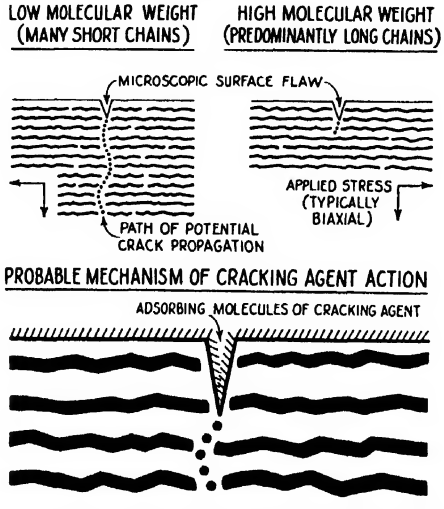


FIG. 10. Simplified scheme of possible cross section of sheet of biaxially stressed polyethylene (in which one of the stress arrows should be imagined as perpendicular to the page). Crack propagation is shown by dotted lines; electron micrographs of crack replicas indicate that paths may run around crystallites or spherulites. Cross-sectional size of cracking agent molecules should be considered as about same as that of polymer molecules. Replicas have shown that tiny surface flaws are present.

shock-cooled samples.[17,22] The latter condition is particularly important as it arises, often in uncontrolled ways, in technical processing of thermoplastics. However, while the life of these metastable states of order may be very long (years) for polar polymers like the polyamides and cellulose derivatives it is short for weakly-interacting chains like polyethylene. Thus, while spherulites may not be able to grow, and hence quenched samples of polyethylene may stay quite transparent, lateral ordering of the chains occurs rapidly at 25°C.

This is shown by the microphotometer tracings of x-ray diagrams, in Fig. 8. Specimen A was thoroughly quenched, by much more rapid cooling than could occur in practice. The 200 spacing is hardly resolved, and the 110 is diffuse with a half-breadth of 8.7 (in arbitrary units; the half-breadths were obtained after subtracting background corrections). While this sample had a lower yield stress than the same polymer annealed to condition C, its ultimate stress-strain properties were not radically different. Further, after 24 hr. at 25°C, the sample had re-ordered to give the pattern represented by Fig. 8B. The 200 spacing feature is resolved, although weak with a half-breadth of 7.4, while the 110 has intensified and sharpened to 7.7. This change indicates (1) that important physical differences in polyethylenes which have been processed with different thermal history will be lost after a few days standing at room temperature (this refers to lateral ordering, *not* to any long chain axis orientation or "molding strains"); (2) the energy differences between the metastable condition A and more ordered condition B are small, since activation energy for extensive transformation seems to be provided by kT at 25°C. This agrees with the small difference in physical properties between the two states, in contrast to the polar polymer case.[16]

Finally, the curve of Fig. 8C is for a sample which was allowed to solidify slowly, and was held at an elevated temperature for several hours. The 110 feature now has a half-breadth of 6.9, and the 200 of 7.4. This sample had maximum growth of spherulites, as well as presumably roughly[20] equilibrium crystallinity, but the conclusion is that over-all molecular order is not notably more than would eventually be reached by a practical sample standing at room temperature. Also, it was found that a polymer of molecular weight properties in the low elongation-brittle fracture class of Figs. 1 and 2 was not converted to a tougher, cold-drawing behavior by extreme quenching, and testing after standing a day. No doubt a polyethylene just on the verge of being able to orient under complex stressing, in the annealed condition could be oriented after thorough quenching. But these circumstances do not often arise; curves 8A and 8C represent practically the whole range of metastable states of the present type of commercial polyethylene. Detailed studies of polymers covering the 8A-8C extremes showed biaxial stress-strain curves of the latter lying about 240 p.s.i. above the disordered samples. Moduli from the first, linear portions of the curves were 13,200 p.s.i. for the quenched and 34,800 p.s.i. for the annealed polymers. The complete stress-strain curves appear in Fig. 9.

Stable reductions in degree of crystallinity at elevated temperatures are currently being studied in relation to complex stressing at those temperatures. Presumably near 100°C there will be easier and more isotropic crystallite orientation than at 24°C.[18] The over-all reduction in crystallites also shows up in stress-strain properties.[23] Hence, many of the effects in this report may be changed at other temperatures.

CRACKING AGENTS AND COMPLEX STRESSING

Crazing or micro-cracking of polymers has been long studied, especially in technical development of rigid

[20] R. B. Richards, Trans. Faraday Soc. 41, 127 (1945).
[21] Raine, Richards, and Ryder, Trans. Faraday Soc. 41, 56 (1945).
[22] Fuller, Baker, and Pape, J. Am. Chem. Soc. 62, 3275 (1940).
[23] Stein, Krimm, and Tobolsky, Textile Research J. 19, 8 (1949).

polymers such as polystyrene and polymethyl methacrylate, and in many finishes. It arises from shrinkage stresses when volatile components (such as unreacted monomers) escape, from external stresses,[24] and from internal stresses reflecting inhomogeneous molecular packing.[25] Spectacular acceleration of crazing occurs with application of many organic liquids or their vapors to the surface of these polymers, under either internal or external stress.[24,25] Stressed polyethylene (of certain molecular weight range) also shows this brittle cracking, by "mobile polar liquids."[8] The present work has revealed that complex stresses are strikingly effective in this surface-agent attack; biaxial tension causes extensive cracking in the presence of alcohols, esters, organic salts, etc., many times faster than uniaxial tension. However, as with the effects of biaxial tension alone the effects with surface-agents are also sensitively affected by molecular weight.

Cracking liquids have been called "surface-agents" above, following a mechanism postulated for crazing action, in connection with earlier work.[25] This is shown schematically in Fig. 10, where the "lyophilic" or non-polar parts of the adsorbed molecules of cracking liquid are assumed to penetrate the outermost layers of polyethylene chains. These adsorbed molecules presumably tend to spread to form Langmuir-type films. As they cover the polymer surface, they run along into tiny flaws and cracks and, at the apex of these cracks, probably exert a film spreading pressure. This film pressure, along with weakening of the polymer mole cohesion near the apex by swelling by the non-polar portions of the sorbed molecule, may add to an external stress to yield crack propagation. Biaxial external stresses, whatever the reason, cause especially rapid cracking in the presence of such agents.

Likewise, solvents like benzene, toluene, etc., which can actually swell polyethylene, even at room temperature, cause brittle fracture under biaxial stress. This seems to be just a weakening of the intrinsic structure, however, perhaps because of the drastic lowering of number average molecular weight in such a "plasticized" system. (Before complete penetration, these solvents presumably act as surface agents.)

[24] W. B. Klemperer, Applied Mechanics, T. von Kármán Anniversary Volume, 328 (1941).
[25] W. O. Baker, U. S. Patent 2,373,093 (1945).

GENERAL SIGNIFICANCE

Reactions of other microcrystalline polymers to complex stresses is being explored. Many queer cases of brittle failures in use by polymers normally considered (by uniaxial criteria) to be tough may have come from biaxial stressing in service. Certain brittle point tests (the test for incidence of brittle fracture, usually in rubbery materials at low temperatures) also seem to introduce varying amounts of biaxial tension in the bent sample. This may be one of the poorly controlled elements in brittle point evaluation. Likewise the great sensitivity to cracking agents of polymers already studied under complex stresses suggests that effects of liquid immersion baths in brittle tests may have been underestimated.

Bending of cable sheaths causes mostly biaxial stressing as do many common deformations of tubes, bottles, and the like. Likewise, paints and lacquers on tubular or other surfaces which are bent can become subject to intense biaxial stressing. In these cases, the strains are often small enough so that the given polymer material is commonly thought to have much reserve elongation before rupture, *again based on conventional uniaxial tests.* The present work at least indicates that many of these usual engineering practices must be re-examined. In the case of cable sheath, a complete upward revision of ideas of acceptable molecular weight polymer was necessary because of effects of complex stressing when the cable was bent and twisted.

Frozen, crystallized natural rubber is being studied to see if rubber crystallites not *formed* in the direction of stress show blocking. This would be a variation of the normal behavior found for initially amorphous rubber biaxially extended up through its presumed stress-crystallization point.[6]

Several technical tests such as paper bursting tests operate under complex stresses. Microcrystalline polymer films, especially those of interest because of low moisture permeability, such as polyethylene, rubber hydrochloride, polytetrafluorethylene, and polytrifluoro-monochlorethylene may show special properties in these tests.

ACKNOWLEDGMENTS

Mr. N. R. Pape has contributed broadly to these x-ray and mechanical studies. We wish also to thank Dr. R. D. Mindlin and Mr. W. T. Read for discussions about the diaphragm and tube tests.

D.C. Characteristics of Silicon and Germanium Point Contact Crystal Rectifiers. Part I. Experimental*

H. J. YEARIAN
Physics Department, Purdue University, Lafayette, Indiana
(Received August 29, 1949)

Typical d.c. current-voltage characteristics obtained for Si and Ge crystal rectifiers are described. A survey of the published theories of the rectifier shows that none of them will account for the principal features of the observed characteristics. The most obvious discrepancy is in the low resistance direction of flow where the logarithm of the current rises with increasing voltage at only a fraction of the rate indicated by the theories. The need for a more flexible theory is pointed out and the conditions which it must meet, both as a function of voltage and temperature are outlined.

THE development of microwave radar has been accompanied by wide use of point contact crystal rectifiers, and a revived interest in the details of their behavior. Since a complete understanding of their operation at high frequencies must be predicated upon their low frequency properties, extensive investigations of their d.c. characteristics have been made.

It is the purpose of Part I to give typical d.c. characteristic data of silicon and germanium rectifiers, the two types of interest at present in the microwave region,** and to show that the published theories of the rectifier do not account for the important features of these characteristics. Part II† will discuss an extension of the diode theory, the so-called multicontact theory, which is in agreement with the experiments.

The usual rectifier for high frequency applications consists of a metal point of approximately 10μ-diameter, in contact with a suitably alloyed crystal of silicon or germanium semi-conductor. The ratio of current at a given applied voltage in the low resistance, or forward, direction to that in the high resistance, or backward, direction, may be several hundred. The silicon now in use is a hole, or P type, conductor for which the direction of easy flow of electrons across the rectifying junction is from metal to semi-conductor. The polarity is opposite for germanium, since this material is used as an electron, or N type, conductor.

In series with the variable resistance of the rectifying barrier is the resistance encountered as the current spreads outward from the contact into the semi-conductor. This "spreading resistance" has a value, $R_s=\rho/(2d)$, where ρ is the resistivity of the semi-conductor and d is the diameter of the contact. All experimental data must be corrected for the potential drop in this resistance before comparison with theories of the rectifying barrier.

A number of theories of the rectifying contact have been developed[1] which differ in their treatment of the problem according to the assumptions made concerning the nature of the rectifying barrier or blocking layer.

The barrier may be assumed to be an artificial insulating layer on the surface of the semi-conductor, or a depletion layer formed near the surface by absence of the impurity centers which produce the conductivity of the semi-conductor. It may also be assumed to be the natural barrier formed when direct contact is made between a semi-conductor and a metal having a different work function.

The electron potential energy diagrams for these cases are shown schematically in Fig. 1, for an N type semi-conductor. Analogous diagrams may be made for the P type.

For the usual densities of donator states slightly below the bottom of the conduction band of the semi-conductor, the Fermi level lies approximately half way between the donator levels and the bottom of the conduction band. Well within the semi-conductor, when at equilibrium with no voltage applied, this level must lie at the same height as the top of the Fermi lake in the metal. When, therefore, the work function of the semi-conductor is less than that of the metal,†† a potential barrier is formed of height equal to the difference of the work functions, the contact potential, V_c.‡ With an insulating layer on the surface of the semi-

* This work was performed under OSRD Contract OEMsr-362 and was reported in an NDRC Report dated December 3, 1942, PB5189.

** The "High Back Voltage" type of germanium rectifiers have high frequency properties which are significantly different from the microwave type, but the low voltage d.c. characteristics are similar and most of the conclusions of this paper will apply. See reference 5, and NDRC Reports Purdue University: 14-341, PB5201, Nov. 1, 1944; 14-375, PB5202, Dec. 26, 1944; 14-413, PB5204, March 19, 1945.

† Johnson, Smith, and Yearian (to be published).

[1] W. Schottky, Zeits. f. Physik **118**, 539 (1941); N. F. Mott and R. W. Gurney *Electronic Processes in Ionic Crystals* (Oxford University Press, New York, 1940); W. Schottky and E. Spenke, Ver. Siemen-Werk. XVIII 31-68 August 1939; W. Schottky, Zeits. f. Physik **113**, 367 (1939); N. F. Mott, Proc. Roy. Soc. **A117** 27 (1939); Proc. Camb. Phil. Soc. **34**, 568 (1938); B. Davydow, Tech. Phys. U.S.S.R. **5**, 87 (1938); A. H. Wilson, Proc. Roy. Soc. **A136**, 487 (1932); J. Frenkel and A. Joffe, Physik. Zeits. Sowjetunion **1**, 60 (1932), Lothar Nordheim, Zeits. f. Physik **75**, 434 (1932); C. Wagner, Physik. Zeits. **32**, 641 (1931).

†† These are the work functions under existing conditions, not necessarily those for clean surfaces.

‡ J. Bardeen has investigated the effect of energy states on the surface of the semi-conductor (Surface States and Rectification at a Metal Semi-Conductor Contact, Phys. Rev. **71**, 717 (1947)) and has shown that the barrier height may be nearly independent of the metal work function. No major changes in the results of the rectifier theories outlined below will be occasioned by introduction of the surface states if "V_c" and "contact potential" are interpreted as meaning the height of the potential barrier, however formed.

conductor, either due to an artificial layer of an insulator or to a depletion of donators, there results a nearly linear variation of potential across this region. With no insulator nor depleted layer, when contact is made electrons near the surface will go over to the metal, leaving a positive space charge due to ionized donators. The potential fall in this case is quadratic and the barrier has a thickness, D, of approximately

$$D=[\epsilon V_c/(2\pi e\mathcal{N}^+)]^{\frac{1}{2}}, \quad (1)$$

where ϵ is the dielectric constant and $\mathcal{N}^+$ the density of ionized donators.

The theories may be further differentiated according to the assumption of a thin, or a relatively thick barrier layer.

For the case of an insulator sufficiently thin that electrons may tunnel through, Wilson,[1] and also Mott[1] have shown that if the current and the voltage across the contact are taken positive for electrons flowing from semi-conductor to metal, the net current density is

$$j=-4\pi me(kT)^2ph^{-3} \times\exp(-V_c/kT)[\exp(-eV/kT)-1], \quad (2)$$

where p is the transmission coefficient of the barrier, V is the voltage across the barrier (applied voltage V_a minus the potential drop in the spreading resistance) and the other constants have their usual significance.

As has been pointed out by Mott, this result is contrary to experiment since for an N type semi-conductor it gives the low resistance direction as corresponding to the flow of electrons from metal to SC $(-V)$ whereas, experimentally, this is the high resistance direction. Tunneling effects cannot, therefore, be the fundamental source of the rectification. Such effects are of importance, however, in the detailed analysis of the characteristic in the backward direction and will be discussed below.

In the case of a depletion layer for which the thickness, D, is large compared with the electron mean free path, so that conditions of diffusion exist in the barrier region, Mott[1] finds for the current density

$$j=ebn_0(V_c-V)D^{-1}[\exp(eV/kT)-1], \quad (3)$$

where b is the electron mobility in the semi-conductor and n_0 the electron density at the boundary of the semi-conductor in the absence of current.

In the forward direction for $0<V<V_c$, Eq. (3) is similar in behavior to that for tunneling, but the sign is reversed as experiment requires. The exponential is the dominant term, so that the slope of the curve ln$i$ *versus* V, $V\gg kT/e$, should be only slightly less than e/kT. In the backward direction, $V<0<V_c$, the current should quickly approach an approximately ohmic behavior of differential resistance $D/(ebn_0)$ determined by the first factor. The back current asymptote extrapolated to $V=0$ should give an intercept ebn_0V_c/D, the current in each direction across the barrier at zero applied voltage.

The effect of image force is to lower the barrier and this will have the effect of making the backward current increase more rapidly than indicated by Eqs. (2) or (3).

The theory of the thick natural barrier has been considered in detail by Schottky and collaborators[1] especially for the low voltage region. For semi-conductors of the "exhausted" type, in which the donators are completely ionized, the characteristic may be written in the form

$$j=\sigma_b(8\pi e\mathcal{N}/\epsilon)^{\frac{1}{2}}(V_c-V)^{\frac{1}{2}}[\exp(eV/kT)-1] \times\{1-\exp[-e(V_c-V)/kT]\}^{-1}, \quad (4)$$

where $\mathcal{N}$ is the density of (completely) ionized donators and σ_b is the conductivity of the semi-conductor at the boundary. For $V_c\gg kT/e$ and $V\ll V_c$, the denominator is approximately unity. The square root varies only slowly, so that in the forward direction, V positive, the current increase is nearly exponential with a logarithmic slope of $\approx e/kT$. In the backward direction the current should increase slowly according to the square root.

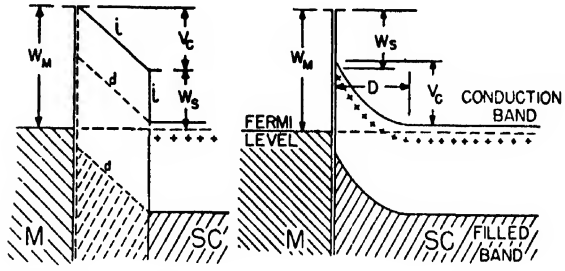


FIG. 1. Potential barrier between metal, M, and N type semi-conductor, SC, at equilibrium with no voltage applied. W_m and W_s, work functions of metal and semi-conductor; D, thickness of natural barrier; $V_c=W_m-W_s$, barrier height; +++ donator levels. Left, insulator (full lines i) between M and SC with potential differences V_c across insulating barrier; left (dotted lines d) depleted layer of SC next metal with potential difference V_c across this layer; right, natural barrier of height V_c. Levels at right of each figure will be moved up or down by amount of positive or negative voltage, respectively, applied across the contact.

When the donators are not completely ionized, the so-called reserve region, Schottky and Spenke[1] have shown that at low voltages the factor corresponding to the logarithmic slope e/kT becomes $e/(\delta kT)$, where δ may vary up to three as a maximum. They also show that as the higher forward voltage region is approached, the characteristic curves for the reserve and exhausted cases merge. The logarithmic slope as measured at voltages above 0.1–0.2 volt, therefore, should be $\approx e/kT$ in either case. It is likewise shown that in this region there is no essential difference between the theories for depletion and non-depletion of donators. This is in agreement with the similarity of Eqs. (3) and (4) in the forward direction.

If the natural barrier is thin compared with the electron mean free path in the semi-conductor an electron makes no collisions in the barrier region, and the problem may be treated as that of a diode. This theory has

TABLE I. Electrical properties, mean free path and barrier thickness in several samples of germanium and silicon.

	Resistivity ohm cm	Hall coefficient cm^3 $clmb^{-1}$	n $10^{14}cm^{-3}$	N $10^{14}cm^{-3}$	x_0 10^{-7}cm	D 10^{-7}cm	1 10^{-7}cm	ΔE e.v.
N type Ge	0.0665	280	26.3	26.3	41.5	141	273	0.051
	0.044	61	121	125	19.5	64.5	86	0.039
	0.0194	19.8	372	374	11.1	37.8	64	0.017
	0.00715	7.4	983	983	6.75	23.0	65	*
	0.00500	3.43	2150	2170	4.60	15.7	43	*
	0.0044	2.93	2510	2630	4.23	14.4	41	*
N type Si	0.079	104	70.9	76.8	25.2	85.7	82	0.047
	0.0535	77	95.8	105	21.9	74.5	79	0.072
	0.026	31	238	300	13.9	47.3	74	0.037
	0.0105	12.8	575	624	8.88	30.2	51	0.032
	0.0104	10.4	709	792	8.02	27.3	62	0.030
	0.0101	9.8	784	819	7.78	26.5	61	0.020
	0.0075	7.9	934	980	7.00	23.8	65	0.024
P type Si	0.048	5.4	1370	3400	9.0	30.6	7.9	0.081
	0.048	4.2	1750	4400	7.9	26.9	6.2	0.077
	0.029	1.74	4230	22,200	5.1	17.3	4.2	0.084

* Hall curve too flat to permit reliable calculation of ΔE.

been developed by Bethe[2] and Sachs[3] and gives the result

$$j=j_0\exp(-eV_c/kT)[\exp(eV/kT)-1], \qquad (5)$$

where j_0 is the thermal current density in the semi-conductor, $j_0=ne[kT/2\pi m)]^{\frac{1}{2}}$ and n is the electron density in the conduction band. The terms of Eq. (5) outside the brackets represent the electron current over the top of the barrier from metal to semi-conductor and with no voltage applied, the equal current in the opposite direction. As the voltage is varied the first current is constant (without image force or tunneling effects) while that from semi-conductor to metal varies exponentially with voltage. This theory therefore also gives a logarithmic slope in the forward direction of e/kT, while the current density in the backward direction quickly increases to a saturation value $j_0\exp(-eV_c/kT)$.

The intermediate case of a natural barrier which is comparable in thickness with the mean free path has been investigated by Herzfeld[4] to show that the current will differ from that of the pure diode theory, Eq. (5), by less than five percent in the range ± 1 volt.

The forward characteristics predicted by all of these theories are much the same and the current may be approximately represented by

$$i=i_0[\exp(eV/kT)-1], \qquad (6)$$

where i_0 may vary slowly with voltage and is strongly dependent on temperature and contact potential.

The backward characteristics range from the saturation type of Eq. (5) to the approximately ohmic type of Eq. (3). These will be modified in the direction of increased currents by image force and tunneling effects.

A decision between the thick or thin natural barrier, and between exhausted or reserve region theories can be made from measurements of the Hall coefficient, R, and resistivity, ρ, as functions of temperature. From these data one may calculate the mean free path 1; the depth, ΔE, of the donator levels below the bottom of the conduction band or of acceptor levels above the top of the full band in the case of N and P semi-

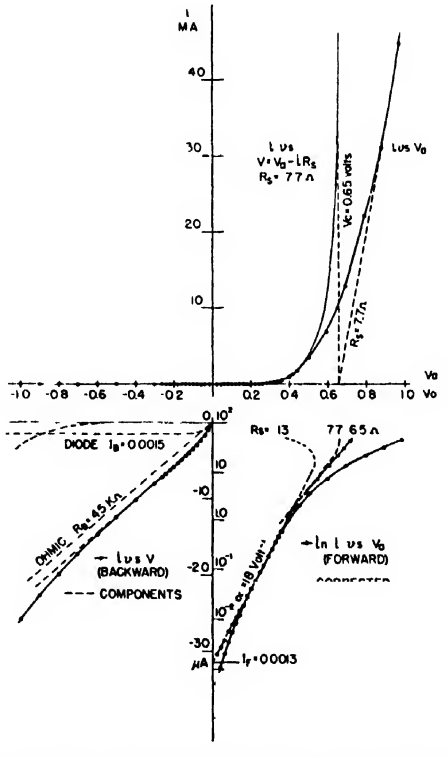


FIG. 2. Typical rectifier characteristic and analysis (W point on Ge, 300°K). Above, linear plot. Below, logarithmic plot of forward and enlarged linear plot of backward characteristic, showing method of correcting the former for spreading resistance and the analysis of the latter into its components.

[2] H. A. Bethe, NDRC Division 14 Report 43-12, MIT November 23, 1942.

[3] R. G. Sachs, NDRC Report 14-129, PB5190 Purdue, September 10, 1942 and PB5195 June 15, 1943.

[4] K. F. Herzfeld, NDRC Division 14 Report, PB5199 (May 5, 1944).

TABLE II. Electrical parameters of Ge and Si rectifiers.

Sample	Type	T°K	α volt^{-1}	$s=\alpha kT/e$	I_F ma	I_B ma	R_B 10^3 ohms
5Id-4	Ge-W	243	22	0.46	3×10^{-4}	4×10^{-4}	254
		295	20	0.52	3×10^{-3}	3.2×10^{-3}	77
		394	8.9–6.2	0.30–0.21	0.25–4.0	0.24	5.8
		549	7.0–5.0	0.33–0.24	2.0–3.0	0.3–0.5	0.09
5Id-2	Ge-W	286	7.4	0.18	0.20	0.25	1.1
		436	6.3	0.24	1.0	0.35	2.0
		506	4.9–2.8	0.21–0.12	2.0–4.0	0.60	0.2
Ge-35	Ge-Pt Ir	165	17.0	0.24	1.3×10^{-4}	$<5\times10^{-5}$	$>1.2\times10^{3}$
		239	16.6	0.34	7.0×10^{-4}	$<1\times10^{-4}$	$>1.6\times10^{3}$
		250	17.0	0.36	9.0×10^{-4}	1×10^{-1}	91
		269	18.0	0.41	1.1×10^{-3}	4.0×10^{-4}	78
		298	16.4	0.42	3.5×10^{-3}	1.6×10^{-3}	39
		299	17.5	0.45	3.0×10^{-3}	1.5×10^{-3}	46
		333	15.9	0.45	8.0×10^{-3}	6.0×10^{-3}	25
		351	17.0	0.51	8.0×10^{-3}	4.7×10^{-3}	31
		360	15.5	0.48	1.5×10^{-2}	1.2×10^{-2}	18
		393	15.3	0.52	3.0×10^{-2}	2.1×10^{-2}	13
GE-40	Ge-Pt Ir	167	19.2	0.28	2.0×10^{-5}	$<5\times10^{-5}$	>300
		241	16.2	0.33	5.0×10^{-4}	$<5\times10^{-5}$	>100
		251	17.7	0.38	3.0×10^{-4}	$<5\times10^{-5}$	>100
		271	18.2	0.42	4.0×10^{-4}	$<1\times10^{-4}$	167
		298	18.3	0.47	7.0×10^{-4}	$<1\times10^{-4}$	120
		299	19.1	0.49	6.0×10^{-4}	$<2\times10^{-4}$	82
		299	18.3	0.47	7.0×10^{-4}	$<1\times10^{-4}$	100
		333	17.7	0.50	2.0×10^{-3}	5.0×10^{-4}	74
		351	17.9	0.54	3.3×10^{-3}	1.1×10^{-3}	46
		360	17.3	0.53	4.4×10^{-3}	2.0×10^{-3}	53
		394	15.8	0.53	1.5×10^{-2}	7.5×10^{-3}	24
9 Ha	Ge-W	298	7.8	0.20	0.80	0.013	32
		358	11.9	0.36	0.71	0.28	5.0
		391	11.5	0.38	0.95	0.87	1.7
		426	9.0	0.33	2.5	3.0	0.50
		468	6.6	0.27	5.5	7.1	0.15
		521	7.0	0.31	10.2	7.8	1.9×10^{-3}
		574	4.4	0.22	22		7.0×10^{-3}
267	Ge-Pt Ir	298	14.4	0.37	6.0×10^{-3}	2.2×10^{-3}	16
172	Ge-Pt Ir	298	17.7	0.45	7.0×10^{-5}	$<1\times10^{-5}$	$>2\times10^{3}$
462	Ge-Pt Ir	298	34.7	0.88	7.0×10^{-5}	$<1\times10^{-5}$	$>2\times10^{3}$
2A-1′	Ge-W	298	10.6	0.27	2.4×10^{-2}	3.0×10^{-3}	7.2
2A-2′	Ge-W	298	14.6	0.37	2.0×10^{-2}	2.0×10^{-3}	>8.3
2A-3′	Ge-W	298	13.9	0.35	7.4×10^{-3}	$<2\times10^{-3}$	>11.9
2A-4′	Ge-W	298	11.7	0.30	5.0×10^{-2}	4.0×10^{-3}	3.2
2A-5′	Ge-W	298	10.7	0.27	3.0×10^{-2}	3.0×10^{-3}	4.6
2A-2	Ge-W	298	10.4	0.27	1.1×10^{-2}	2.5×10^{-4}	11.2
2A-3	Ge-W	298	12.0	0.31	3.0×10^{-2}	5.0×10^{-3}	3.0
2A-4	Ge-W	298	11.6	0.30	2.6×10^{-2}	5.0×10^{-3}	3.7
2A-5	Ge-W	298	13.3	0.34	9.0×10^{-3}	3.0×10^{-3}	6.8
2A-6	Ge-W	298	13.1	0.33	2.6×10^{-2}	5.0×10^{-3}	5.7
6Rb-1	Ge-W	298	22.0	0.56	7.0×10^{-3}	2.0×10^{-3}	14.3
6Rb-2	Ge-W	298	14.4	0.37	4.9×10^{-2}	1.4×10^{-2}	2.2
C-4	Si-W	298	18.4	0.47	6.0×10^{-3}	$<1\times10^{-3}$	9.4
C-6	Si-W	298	18.4	0.47	5.0×10^{-3}	1×10^{-3}	18.7
BTH-E1	Si-W	298	21.0	0.54	2.5×10^{-2}	$<5\times10^{-3}$	>3.2
BTH-E2	Si-W	293	16.4	0.41	1.0×10^{-2}	3.0×10^{-3}	7.0
BTH-E3	Si-W	177	11.0	0.17	2.0×10^{-2}	$<2\times10^{-3}$	3.9

conductors, respectively; the electron density, n; the density of donator or acceptor levels, N, and by Eq. (1) the natural barrier thickness, D. These data for several samples of silicon and germanium at 300°K are abstracted in Table I from the work of K. Lark-Horovitz and collaborators.[5] The thickness, D, is calculated for $V_c=0.3$ volt. The distance x_0 over which the potential barrier rises by kT/e is included, since this may be the more nearly correct critical distance to be compared with 1.[2]

In germanium the mean free path is three to ten times x_0, and in silicon approximately equal to x_0. In view of the work of Herzfeld,[4] the diode theory should, therefore, apply to these materials if the rectification is due to a natural barrier.

[5] K. Lark-Horovitz, "Preparation of semi-conductors and development of crystal rectifiers," NDRC Report 14–585, Purdue University.

That an insulator on the surface, or a depletion layer, is necessary for rectification in these materials seems highly improbable, since the general features of their characteristic curves are not at all sensitive to surface treatment. A freshly broken surface of a homogeneous single crystal may give characteristics quite the same as does a suitably polished and etched surface. There are surface treatments in use which could result in a depleted or insulating layer, but these treatments are not essential to the rectification. In any case, characteristics of the type of Eq. (6) are to be expected in at least the forward direction.

Experimental characteristics of typical silicon and germanium point contact rectifiers are shown in Fig. 2. In the forward direction the plot is made semi-logarithmic as well as linear to facilitate comparison with Eq. (6).

The original current data is plotted against the applied voltage V_a, and this voltage corrected to $V=V_a-iR_s$ where R_s is the spreading resistance in series with the contact. Since the effective diameter of the contact cannot be determined with adequate precision to calculate this correction satisfactorily, it is made by trial to give as regular a corrected curve as possible. According to the theories, as represented in Eq. (6), the corrected semi-log plot should be linear for $V \gg kT/e$ ($kT/e=1/40$ volt at 300°K). It is nearly always possible to find a value of R_s (6.5 ohms, Fig. 2) consistent with that calculated from the measured diameter, which will give a linear corrected curve in the usual range of V_a. As is indicated in the figure, however, this linear portion is usually not a continuation of the lower part where the spreading resistance is negligible. The upper slope for normal rectifiers tends to be somewhat less than the slope of the lower part. This is particularly true for germanium. If one attempts to make the upper slope as large as the lower by using a larger value of R_s, (13 ohms, Fig. 2) the corrected curve will usually begin to bend backward to give a negative differential resistance at applied voltages in the neighborhood of one volt. With some R_s between these values one could expect to obtain a corrected curve which approaches a vertical line corresponding to $V=V_c$ as an asymptote. Such a curve is approximated in Fig. 2 with $R_s=7.7$ ohms. This is equivalent to constructing the asymptote of differential resistance 7.7 ohms on the linear plot, cutting the voltage axis at V_c.

To make satisfactorily accurate determinations of R_s and V_c by either of these constructions would require data to higher voltages than is usually attainable because of erratic behavior of the contact. The interpretation of such data is also uncertain due to the strong dependence of the semi-conductor resistivity on temperature. In the case of germanium, the linear plots frequently show no definite region of truly constant slope, but the slope continues to increase as the voltage is raised to values for which it is certain there will be disturbances caused by heating.‡‡ This corresponds in the above analysis to a decrease in the effective R_s and an increase in effective V_c as the applied voltage is increased. In view of these difficulties, perhaps the best one can say at present is that the straight line correction on the semi-log plot gives a minimum value of R_s, and that which produces a negative differential resistance gives a maximum value appropriate to the voltage range used. These uncertainties at the upper part of the characteristic do not affect it appreciably below 0.4–0.5 volts, however, so that comparison with theory can be made in this range.

According to Eq. (6), if a current i_0 is added to the observed forward current, a straight line of slope e/kT should be obtained on a semi-logarithmic plot. Usually it is possible, as is indicated in Fig. 2, to find such a current which will yield a linear region of considerable length. Let us denote this current which corresponds to i_0, by I_F in order to distinguish it from a similar current, I_B, obtained from the backward characteristic. Also denote the slope of the corrected curve by α.

The curves of Fig. 2 are quite typical of silicon and germanium rectifiers. While linear regions of ln$i$ extending over five powers of ten in current have been found, a range of one hundred to one thousand is more common. Characteristics are found which deviate from this linear behavior, usually in the sense of two oppositely curved portions, convex upward at the low currents and convex downward at the higher currents, but these are the exception.

These characteristics are in qualitative agreement with the prediction of theory, but the quantitative agreement is very poor. The result common to all theories discussed above is that the logarithmic slope should be very approximately $\alpha=e/kT=40$ volts^{-1} at 300°K. The experimental slopes are always less than this. The value shown in Fig. 2 of approximately one half e/kT is typical as is evident from the data listed in Table II. Occasional higher values of α do occur, but the one of .88 e/kT quoted for sample 462 is exceedingly rare. Much less uncommon are values down to one-fifth or one-sixth e/kT. Use of Pt-Ir alloy for the metal in contact with germanium has some tendency to increase the value of α as compared with that obtained with tungsten.

It is evident that the approach to the linear portion of the curve is much less rapid than is indicated by the theories, so that the data can usually be represented by a characteristic of the form $i=I_F[\exp(\alpha V)-1]$.

The wide range of I_F indicated in the table is not inconsistent with the diode theory. One may set $I_F=i_0.1=Aj_0\exp(-eV_c/kT)$ where A is the area of the contact. Using 10^{-6} cm^2 as an average value of A and with $j_0\simeq 10^6$ amp/cm^2 computed from the Hall

‡‡ This apparent inconstancy of spreading resistance is under investigation, particularly for the "high back voltage" type rectifier. See reference 5; Ralph Bray, Phys. Rev. 74, 1218 (1948).

coefficient measurements, one finds a range of I_F from 1 ma to 0.2μa corresponds to a variation of V_c between 0.17 and 0.39 volt. These are reasonable limits, not in disagreement with the estimates to be made from the spreading resistance correction discussed above.

Granting, however, the I_F is correct, it follows from the exponential dependence of the current on α, that with α one-half e/kT the current at 0.3 volts is $\frac{1}{4}$ percent of the theoretical value, and for α one-third e/kT, it is only 1/30 percent of theoretical. Clearly some modification of the theories is necessary.

A similar disagreement has long been known in the low voltage region; e.g., Wagner, Schottky and Spenke[1]. Although for a reserve semi-conductor one has flexibility in α (Schottky α_1) by the factor of $\frac{1}{3}$, values for Cu_2O, MoS, TeO_2, etc., are found as low as $\frac{1}{8}e/kT$. It may be possible in the low voltage region to account for the remaining discrepancies by the assumption of essentially ohmic imperfections in the barrier. In the higher voltage region, this solution is not possible, since the minimum resistance which can be assigned from the back characteristic will contribute nearly insignificant currents to the forward characteristic at voltages above about 0.1 volt. In the rather favorable example of Fig. 2, if currents equivalent to the entire back current are subtracted from the forward characteristic, the value of α is increased from 18 to only 20.4 volts^{-1}.

The back direction characteristic in Fig. 2 has been plotted on an enlarged linear scale. It is evident that the current does not saturate as required by the simple diode theory, but in general it may be analyzed into a diode component, an ohmic component, and a rapidly increasing component or "tail." The latter part is frequently nearly exponential in character and may, therefore, be ascribed to tunneling through a thin barrier, as suggested by the reversed rectification in this case, Eq. (2), or it may be due to the effect of image force in lowering the barrier. These processes will be discussed below.

The diode-like component, I_B, and the resistance of the ohmic component, R_B, are listed in Table II. It is evident, contrary to the diode theory, that I_B is not the same as I_F. It may be larger or smaller, but is usually smaller, particularly at low temperature. Determination of a new I_F from the forward characteristic after correcting for the ohmic contribution to it, as discussed above, does not in general bring the two currents into coincidence.

The existence of the important ohmic component of the back current suggests that the barrier is of the thick depleted layer type leading to Eq. (3). From the data given and the discussion above, however, it seems highly unlikely that this is necessarily the case. It would rather seem that irregularities which have an essentially ohmic behavior must exist in the thin natural barrier.

The diode theory, with the assumption of such ohmic regions, combined with some form of tunneling and image force barrier lowering should give agreement with the back direction characteristic. Introduction of such modifications also affects the forward characteristic. The ohmic regions themselves will not appreciably reduce the discrepancy between the theoretical and experimental logarithmic slope, as was shown above, but the other effects might be expected to do so.

The effect of barrier lowering by image force has been investigated by Bethe[2] and others[3, 6] and the tunnel effect by Courant.[6] The effect of image force is to lower the barrier by an increasing amount as the voltage is reduced in the forward direction or increased in the backward direction. The tunnel effect may be expressed as an additional lowering.

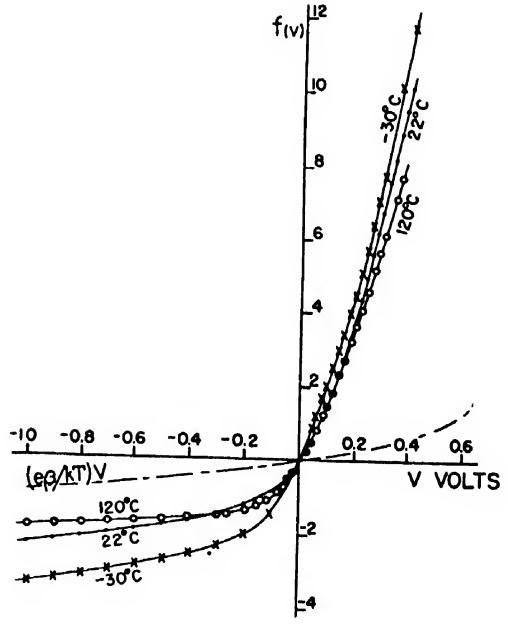


FIG. 3. Typical behavior of the exponent $f(V)$ in a corrective term, $\exp[-f(V)]$, to bring the simple diode theory into agreement with experiment (Data of 5Id-4, Table II). Dashed curve shows the type of correction afforded by the barrier lowering modification of the theory.

The total potential lowering from the original value V_c, may be expressed as $\Delta V_c=\beta(V_c-V)$ where β[Courant $(1-\theta)$] is positive and nearly constant for $V\ll V_c$, and increases slowly as V approaches V_c. Subtracting this correction from the V_c of Eq. (5) gives

$$j=j_0\exp[-e(1-\beta)V_c/kT]\exp(-e\beta V/kT)\times[\exp(eV/kT)-1] \quad (7)$$

instead of Eq. (5). For comparison with experiment,

[6] E. D. Courant, NDRC Division 14 report MIT May 17, 1943. See also Phys. Rev. 69, 684 (1946). In this abstract the sign of the corrective exponent should be positive. The impression which might be given by the abstract that there is agreement between the corrected theory and experiment, is misleading. The full paper shows that the essential discrepancy still exists, as discussed herein.

instead of Eq. (6) one should then use

$$i=i_0' \exp(e\beta V/kT)[\exp(eV/kT)-1], \qquad (8)$$

where i_0' is nearly constant except for $V\simeq V_c$.

As was mentioned earlier, this additional factor is of the correct form to agree with the observed "tail" in the back direction. For negative V the variation of the term in brackets is negligible and the back current increases exponentially. As nearly as one can divorce this effect from the diode and ohmic parts, the experimental data require values of β of approximately 0.01 to 0.05. It seems possible to obtain values of this order from the theory.

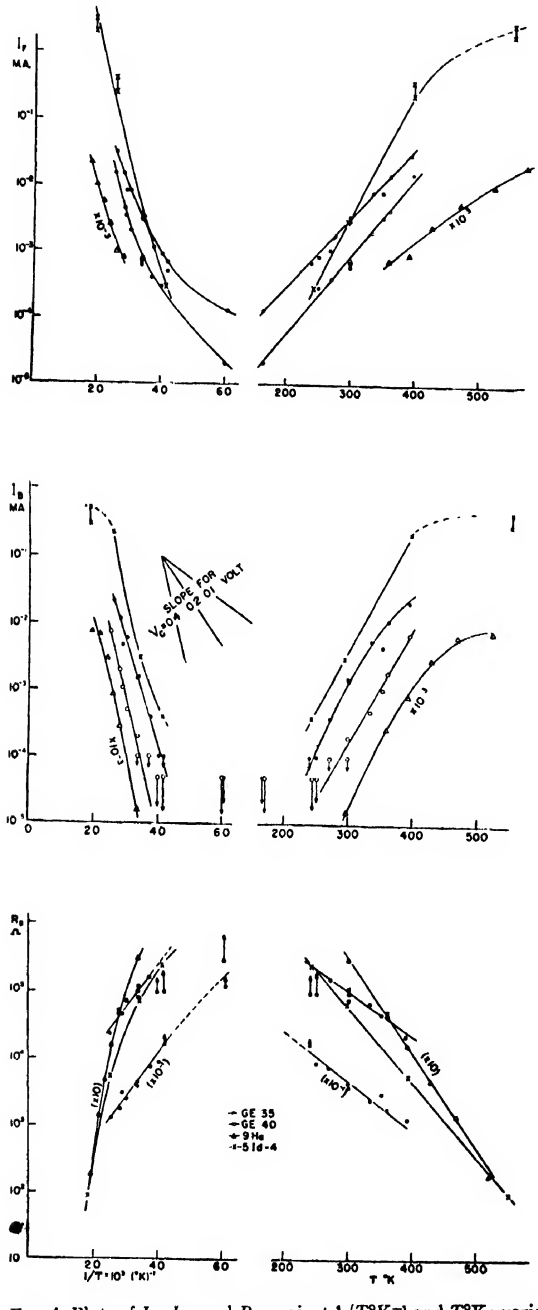


FIG. 4. Plots of I_F, I_B, and R_B against $1/T°K^{-1}$ and $T°K$; variation with reciprocal of temperature is not always linear and slopes of I_F and I_B are not equal.

In the forward direction there would also seem to be an improvement over the simple theory. On expanding Eq. (8)

$$i=i_0'\{\exp[e(1-\beta)V/kT]-\exp[-e\beta V/kT]\} \qquad (9)$$

and since for $V\gg e/kT$ the second term is negligible, one obtains an exponential characteristic whose logarithmic slope has been reduced from e/kT to $(1-\beta)e/kT$.

The reduction possible by this correction cannot, however, be made sufficient to agree with experiment. The values of α usually found range downward from approximately $0.5e/kT$. This requires values of β of 0.5 and larger. For $\beta=0.5$ the predicted value of $\alpha=(1-\beta)e/kT=0.5e/kT$, but in this case Eq. (9) shows that the backward characteristic, V negative, is identical with the forward characteristic and there would be no true rectification. The required values of β larger than 0.5 would give a reversal of sign of rectification. Moreover, quantitative estimates¶ of the magnitude of β show that it cannot be as large as 0.5, except for very low values of V_c.

To be in agreement with experiment, in the terminology of Eq. (8), β would have to increase from $\approx$0.1–0.05 for negative V, to $\approx$0.5–0.7 or more for positive V. While the theories give a variation in this sense, it is negligible in amount compared with the required variation.

The experimental data may be used to determine an empirical function, $-f(V)$, to replace $-e\beta V/kT$ in the corrective term of Eq. (8). The function can be determined only to within an additive constant, because of the uncertainties in the area of contact and contact potential included in i_0'. The curves of Fig. 3, for a germanium rectifier at three different temperatures, have, therefore, been arbitrarily displaced to pass through the origin. In contrast with this rapidly varying function required by experiment, the simple diode theory gives a constant value for $f(V)$ and the barrier lowering theories give a curve which slowly increases in slope as the voltage is increased toward V_c.

The characteristics of a number of crystals have been obtained at various temperatures. Considerable difficulty is encountered in maintaining the same contact over a wide range of temperatures. Some of the more reliable data are listed in Table II.

At the high temperatures, analysis according to Fig. 2 is difficult, because the spreading resistance uncertainty affects a large part of the forward characteristics and also because the measurements in the back direction do not always extend far enough to differentiate between ohmic current and diode current, due to the

¶ References 5 and 6; also V. F. Weisskopf, NDRC Report 14–133 (May 12, 1943).

extremely rapid increase of the latter with temperature. At the low temperatures the back current may become too small for accurate determinations. In these cases ranges of values are given.

The general behavior is that α decreases and I_F and I_B increase very rapidly with increasing temperature, as would be expected from the theories. R_B decreases with increasing temperature, so that whatever mechanism is assumed to provide this ohmic component, it must have an effective resistivity which is very temperature sensitive. The large changes with temperature are demonstrated in the $f(V)$ plots of Fig. 3, remembering that $f(V)$ is the argument of an exponential.

Examination of the values of α in Table II shows that these rectifiers are of different types. For 5Id2, 5Id4, and 9Ha, the ratio of α to the simple theory value, $s=\alpha kT/e$, in general falls with increase of temperature. There is some indication of an initial rise with a maximum at $\approx 400°$K. GE-35 and 40, on the other hand, show a nearly constant value of α, so that the ratio s rises with temperature. Physically, GE-35 and 40 differ from the others in that a condenser discharge had been passed through the contact and the junction was thought to be at least partially welded.||

The variations of I_B, I_F, and R_B with temperature are shown in semi-logarithmic plots as functions of T and $1/T$ in Fig. 4. The apparent disturbance at the highest temperatures may be due partially to physical changes in the contact, although the data were reasonably reproducible. In spite of the limited nature of the data, there seems to be a somewhat better agreement with an equation of the form $\exp(CT)$ than with $\exp(-K/T)$, particularly for I_F, whereas I_F and I_B should each be equal to i_0, and yield a negative exponential in $1/T$, according to Eq. (5). The slope of the log plot against $1/T$ should be $-eV_c/k$ except for possible variation in n, and the small effect of the $T^{\frac{1}{2}}$ coefficient. (The Hall effect data indicate that n is nearly constant over the range of these measurements.) The order of magnitude of V_c required by the data is from 0.1 to 0.65 volts. This range obtained from the temperature variation is reasonably consistent with the range of 0.17 to 0.39 volts estimated earlier from I_F itself. There is, however, no detailed correlation. The values of A and n involved in i_0 are certainly of the same order of magnitude for these samples and a large value of i_0 (I_F or I_B), corresponding to a small V_c should be associated with a small negative slope on the $1/T$ diagram. The lines should tend to radiate from a common point (Ajo) on the $1/T=0$ axis.

From this discussion it becomes evident that a satisfactory theory of the silicon or germanium rectifier must allow the logarithmic slope of the forward characteristic to be reduced below e/kT by a factor which is in itself temperature dependent. The intercepts corresponding to I_F and I_B must not be required to be equal, and must vary approximately exponentially with T. The theory must also provide a mechanism for an ohmic current, as well as a diode component, in the back direction, and the resistance of this component must vary exponentially with temperature, instead of having the relatively slow variation of the semiconductor resistivity.

Part II will be concerned with the development of a diode theory to meet these conditions. The development is based upon a proposal of H. A. Bethe that one abandon consideration of the contact as being a perfectly homogeneous one with all parts having the same contact potential, and adopt the point of view that it consists of more or less isolated regions having lower contact potentials than their surroundings. These regions should be considered to be very small spots ranging in contact potential from very low values up to (or beyond) the values normally expected.

It is essential to consider that the current through any such spot is limited by the local spreading resistance in series with it, or by the fact that the current density through it has a maximum value of j_0. A graphical construction will quickly show that with suitable distributions of the number and size of spots having a given contact potential, one can obtain an α of any value less than e/kT. The temperature dependence of α may be attained by varying the distribution functions. An ohmic component is provided by the contributions of the spots having zero or negative contact potential, for which there is no barrier, and its temperature dependence by the variation with temperature of the number or size of these spots.

|| These experimental crystals were kindly supplied to us by Dr. H. Q. North, General Electric Company, Schenectady, New York.

Electric Breakdown in CO_2 from Low Pressures to the Liquid State*

D. R. Young†
Laboratory for Insulation Research, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received September 26, 1949)

The electric breakdown in CO_2 has been investigated from low pressures through the critical point into the liquid state. Paschen's similarity law is verified for low pressures. At high pressures small departures are observed for long gap-lengths and large departures for small gap-lengths. Simultaneously the scatter of the breakdown voltage becomes independent of illumination and the breakdown strength dependent on the cathode material; this appears due to the onset of field emission. Measurements of prebreakdown currents have yielded values for Townsend's first coefficient as well as for the field emission constants. For small gap-lengths the prebreakdown currents are higher than the normal field emission equation predicts, indicating some new process effective at short gap-lengths. The transition from the gaseous to the liquid state does not produce a discontinuous change of the breakdown voltage.

INTRODUCTION

THREE experimental variables describe the d.c. breakdown characteristics of gases in uniform fields after the gas type, the electrode material, and the source of primary ionization have been chosen: the breakdown voltage v, the gas density ρ, and the electrode spacing d. Paschen established experimentally,[1] that the breakdown voltage of a spark gap is a function of the product ρd only, $v=f(\rho d)$. This equation, in which Paschen used pressure instead of the density, is called the Paschen similarity law. It states that the voltage drop per free path is the important variable.

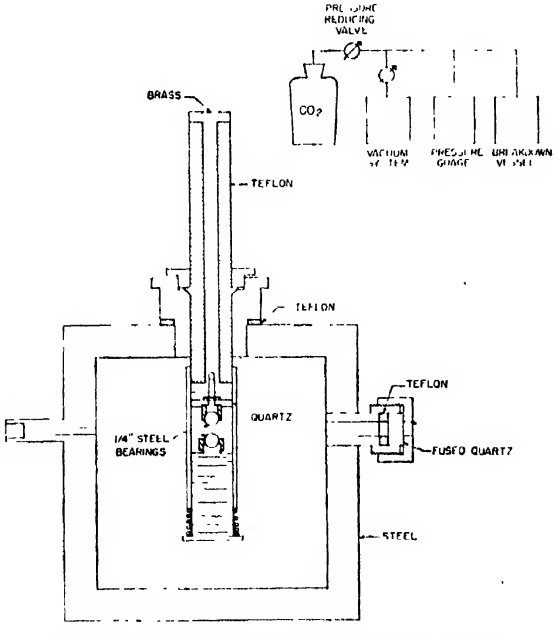


FIG. 1. Schematic diagram of breakdown equipment.

This law cannot hold at high pressures as von Hippel has pointed out.[2] The breakdown field strength in atmospheric air amounts to about 3×10^4 v/cm. For solids and liquids, with the density several thousand times higher, the breakdown strength has risen only to the order of 10^6 v/cm. Hence, instead of about 1.5 v per free path in gases, only 0.01 0.1 v is required for breakdown over the corresponding path lengths in condensed phases. It seems that the breakdown mechanism has changed from the Townsend type, in which the ionization is held in check by electronic excitation processes, to a new mechanism in which the vibration states of the solids and liquids appear to play a decisive role. This latter mechanism has been the object of investigation in this laboratory for some time.[3,4] It is the aim of the research presented here to learn more about

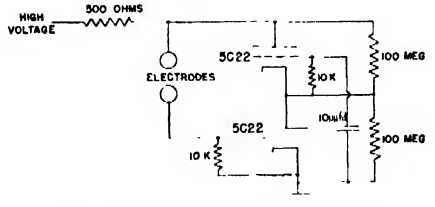


FIG. 2. Electrode protection circuit.

the transition region between the breakdown types in gases and in liquids.

Carbon dioxide was used as the test material since its critical point is readily accessible and its equation of state well known.

EQUIPMENT AND MEASURING TECHNIQUES FOR BREAKDOWN IN THE GASEOUS STATE

A schematic diagram of the apparatus used for the breakdown measurements is shown in Fig. 1. The equipment was evacuated before the CO_2 was admitted. For measurement of pressure a mercury column was used in the lower pressure range (up to two atmospheres)

*This work was sponsored jointly by the ONR, the Army Signal Corps, and the Air Force under ONR Contract N5ori-07801.

† From a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics at the Massachusetts Institute of Technology.

[1] F. Paschen, Wied. Ann. **37**, 69 (1889).

[2] A. von Hippel, Zeits. f. Physik **80**, 19 (1933).

[3] A. von Hippel, J. App. Phys. **8**, 815 (1937).

[4] A. von Hippel and R. S. Alger, Phys. Rev. **76**, 127 (1949).

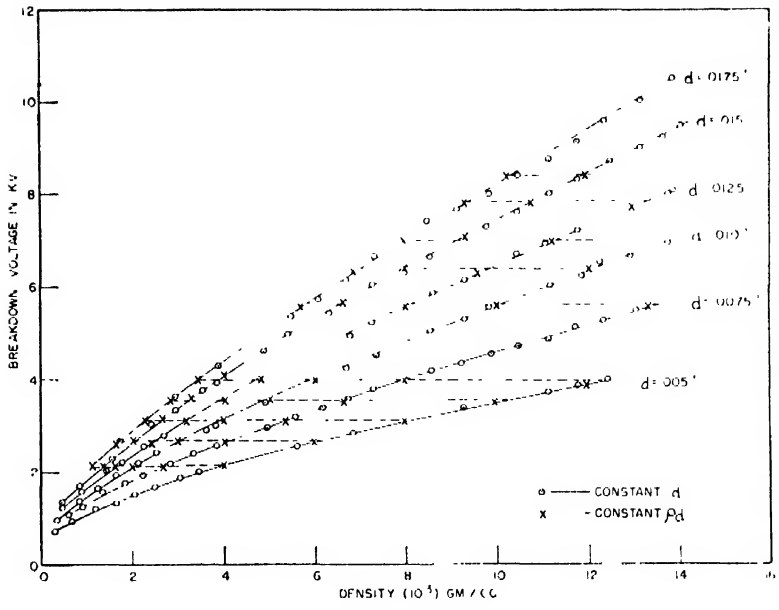


FIG. 3. Breakdown voltage *vs.* density with gap-length as parameter

and Clapp pressure gauges, calibrated by a dead-weight tester, above this limit.

Initially the high voltage insulator of the breakdown vessel was made of polystyrene, but it proved soluble in liquid CO_2 and was therefore replaced by Teflon. Steel ball bearings 0.250 in. diameter served as electrodes since their polish is excellent and their low cost makes them expendable.

Since preliminary experiments indicated that the insulating materials used were not rigid enough to insure an accurate separation if they had to support the pressure, the electrode holder was designed as shown in Fig. 1. The spacing was adjusted by the spring-loaded screw on the lower electrode, and calibration marks 0.0005 in. apart were included. It was possible to adjust the spacing with an accuracy of about ± 0.0002 in.

Teflon was used for the pressure seals with excellent results if the gaskets were entirely confined. A quartz window allowed illumination of the cathode by ultraviolet light.

Without current limitation each breakdown damages the electrode surface and requires its reconditioning. To avoid this damage to the electrodes, the high voltage was removed about 10^{-7} sec. after initiation of the breakdown, by using two hydrogen thyratrons (Sylvania 5C22) in series, triggered by the breakdown current (Fig. 2). Two thyratrons were required since the voltages used exceeded the rating for a single tube.

The high voltage was measured with a multi-range current amplifier used in conjunction with a 10^9-ohm standard resistance. The standard resistance was made up of 100 10^7-ohm one percent resistors which were mounted in 15 copper shells stacked in series. Measure-

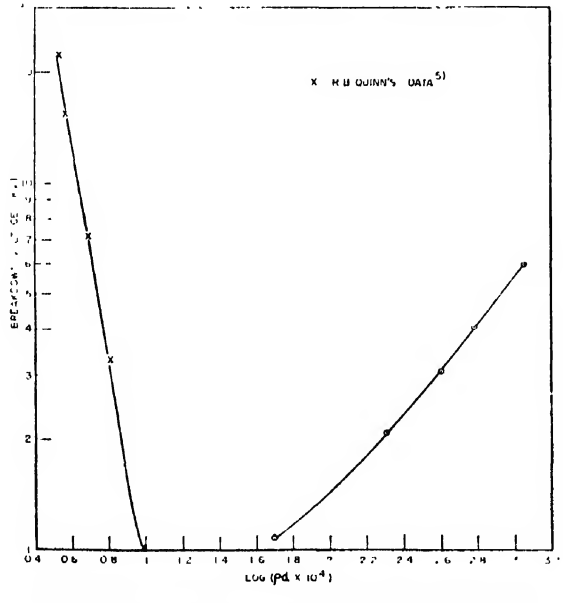


FIG. 4. Breakdown voltage *vs.* ρd (ρ in g/cc, d in mils).

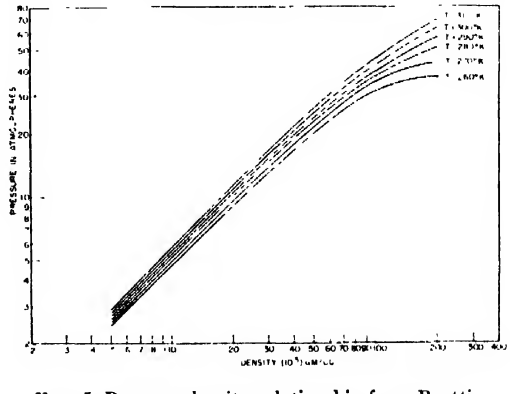


FIG. 5. Pressure-density relationship from Beattie-Bridgman equation.

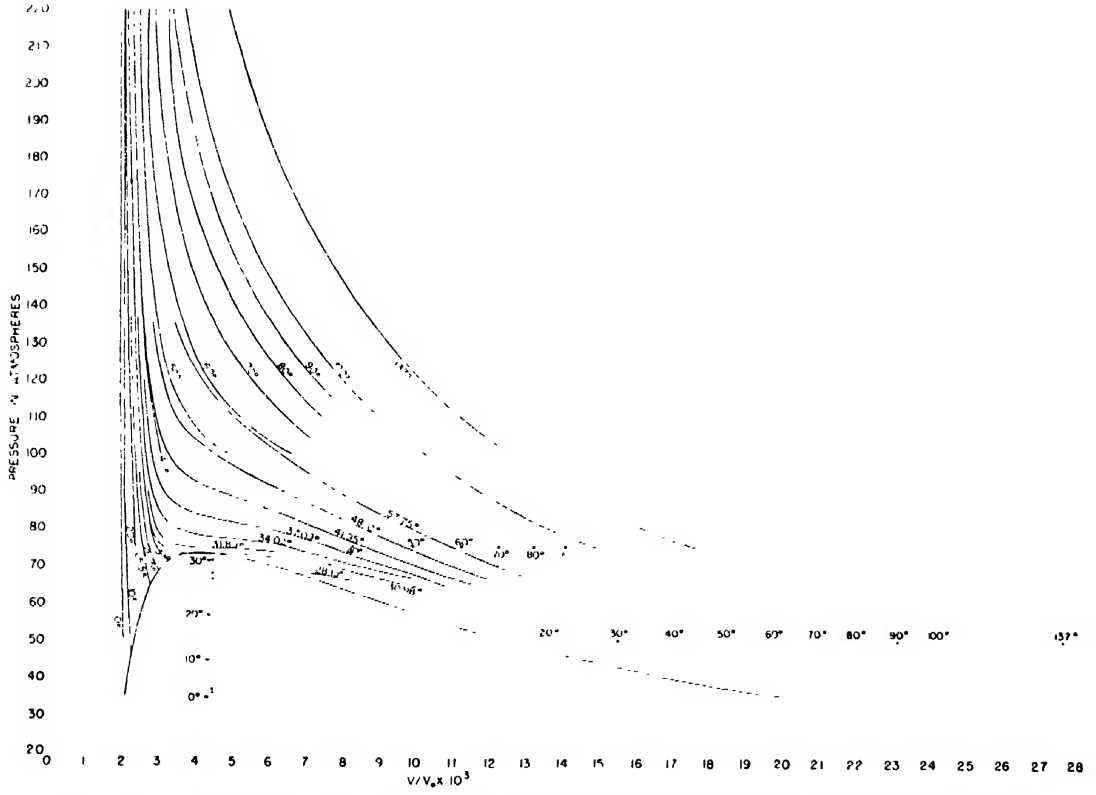


FIG. 6. Pressure-volume data from *International Critical Tables* (see reference 7) in vicinity of critical point

ments of the over-all resistance indicated that it was accurate to within one percent. The shell mounting reduced the field surrounding the resistors and thus prevented corona from this cause.

To reduce the scatter in breakdown voltage due to a scarcity of initiating electrons, the cathode was illuminated by ultraviolet light from a mercury arc. The ratio of electrode diameter to gap-length was kept large enough to assure uniform field geometry.

BREAKDOWN MEASUREMENTS AT LOW GAS PRESSURES

In Fig. 3 the measured breakdown voltage is plotted as a function of gas density with the gap-spacing as parameter. In this region the ideal gas law holds: $P=1.865T\rho$ where P is given in atmospheres, ρ in g/cm, and T in degrees K.

It can be seen that the curves of constant $d\rho$ are horizontal lines, indicating that Paschen's law is obeyed in this region. Consequently, the data of Fig. 3 can be combined into a single characteristic of breakdown voltage *vs.* ρd. Figure 4 shows this curve and its extension to lower pressures taken from the paper of Quinn.[5]

[5] R. B. Quinn, Phys. Rev. **55**, 482 (1939).

BREAKDOWN MEASUREMENTS AT HIGH PRESSURES

For this pressure range the most accurate equation of state is the Beattie-Bridgman equation[6]

$$P=\frac{RT(1-\epsilon)}{V^2}(V+B)-\frac{A}{V^2}, \qquad (1)$$

where $A=A_0(1-a/V)$, $B=B_0(1-b/V)$, and $\epsilon=c/VT^3$,

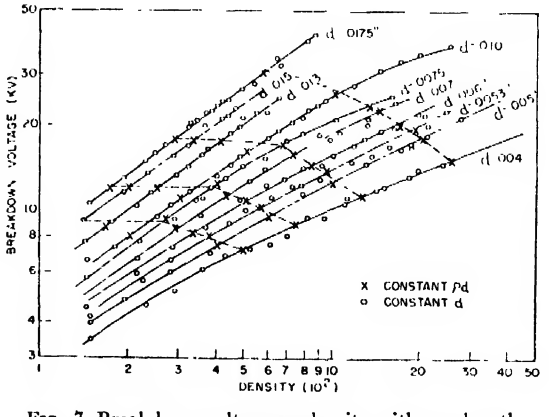


FIG. 7. Breakdown voltage *vs.* density with gap-length as parameter.

[6] J. A. Beattie and O. C. Bridgman, J. Am. Chem. Soc. **49**, 1665 (1927).

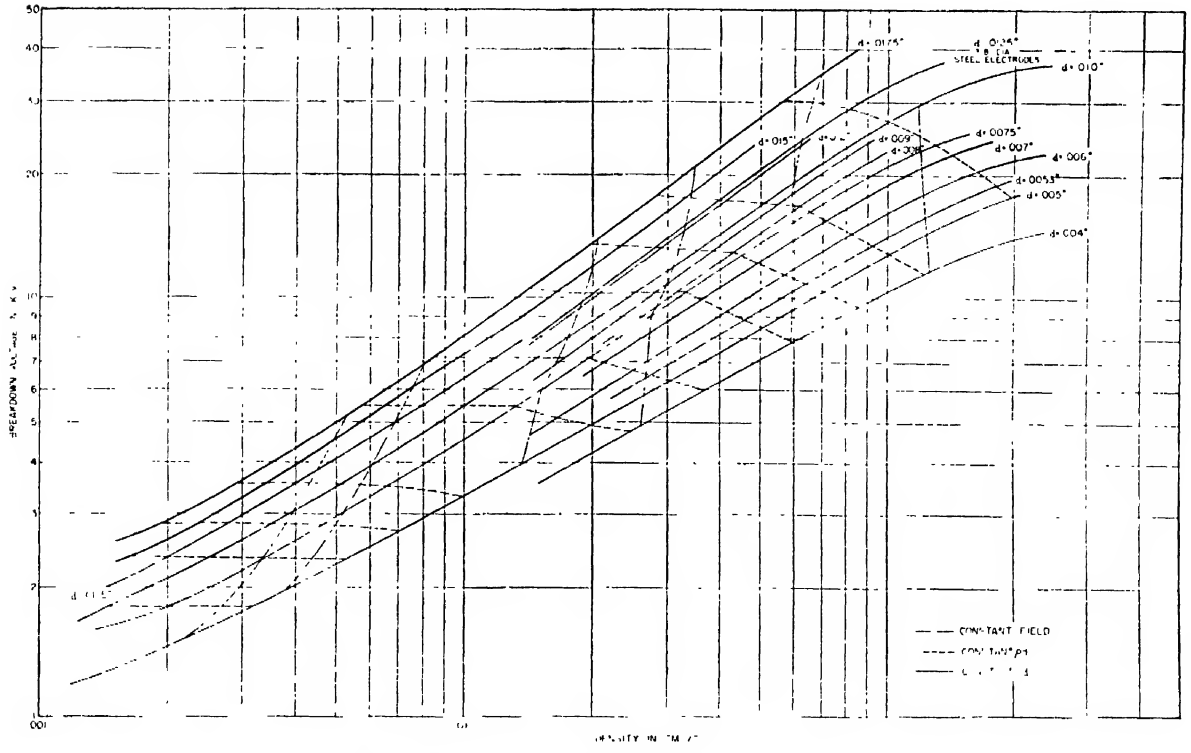


FIG. 8. Breakdown *vs.* density for various gap-lengths.

with P given in atmospheres, V in cc/g, T in degrees K. For CO_2, $R=1.865$, $B_0=2.381$, $b=1.6443$, $A_0=2586.0$, $a=1.6210$, and $C=15\times10^6$. According to Eq. (1) the ideal gas law is valid at room temperature to within one percent for pressures up to three atmospheres. Figure 5 shows the pressure-density relation according to the Beattie-Bridgman equation for various temperatures. In the vicinity of the critical point the relation between pressure and density must be established experimentally. Pertinent data are given in the *International Critical Tables*[7] as plotted in Fig. 6; the relative volume V/V_0 (abscissa) where V_0 refers to the volume at $T=0°C$, and $P=$one atmosphere. The density is correspondingly

$$\rho=\frac{1}{(508V/V_0)}\quad[\text{g/cc}]. \tag{2}$$

The range covered by these measurements extends from densities of 0.01 to 0.24 g/cc (that is, to the density of the vapor above the liquid at room temperature) (Fig. 7).

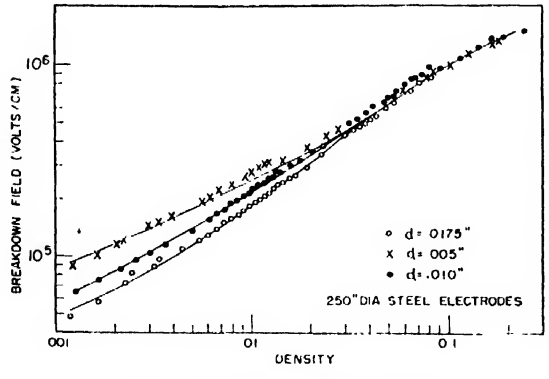


FIG. 9. Breakdown field *vs.* density.

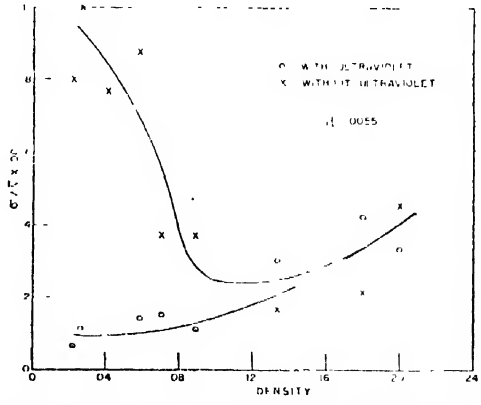


FIG. 10. Effect of illumination on scatter of breakdown voltage measurements.

[7] *International Critical Tables* (McGraw-Hill Book Company, Inc., New York, 1926).

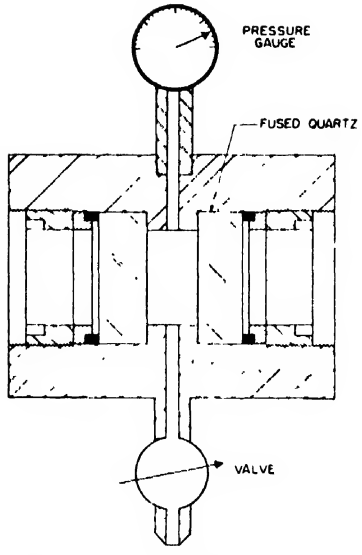


FIG. 11. High pressure absorption cell.

For long gap-lengths small departures from Paschen's law are observed as the deviation from the horizontal of the $d\rho$=constant curves indicates; for small gap-lengths these departures are much larger.

In Fig. 8 the high and low pressure data are combined to show the systematic increase of the deviations. In addition to the ρd curves, lines of constant field strength are indicated. It is seen that for small gap-lengths and high densities these lines are vertical, indicating that the breakdown voltage is determined by the average field strength for a given density, independent of the gap-spacing. If the breakdown field strength is plotted as a function of density, the lines for different gap-spacings consequently converge for high densities (Fig. 9).

If the standard deviation σ is plotted as a function of the density, with the cathode either illuminated or not by ultraviolet light, the curves shown in Fig. 10 are obtained. It is evident that, for the higher densities, the ultraviolet light has no effect.

The ineffectiveness of illumination might be caused by an increase of the ultraviolet absorption with gas density which would prevent the light from reaching the cathode. Since no absorption data were available for CO_2 covering this pressure range, measurements were made using a Cary recording spectrophotometer and the gas chamber shown in Fig. 11. The quartz windows were selected for transparency and had a cut-off limit near 2000A. There was no measurable absorption up to the saturation pressure at room temperature for wave-lengths down to 2000A, which rules out this explanation.

Another possible explanation for the data in Fig. 10 is that field emission provides the necessary initiating electrons. Breakdown measurements were therefore made using zinc-coated cathodes because zinc has a relatively low work function. They were prepared by evaporating zinc on steel ball bearings previously heated to 350°C in a high vacuum in order to obtain the same quality of surface as in the previous experiments. In Fig. 12 the results are compared with those using steel cathodes. At high gas densities the breakdown voltage is appreciably reduced by using zinc-coated electrodes, as would be expected.

At the higher densities the breakdown voltages for

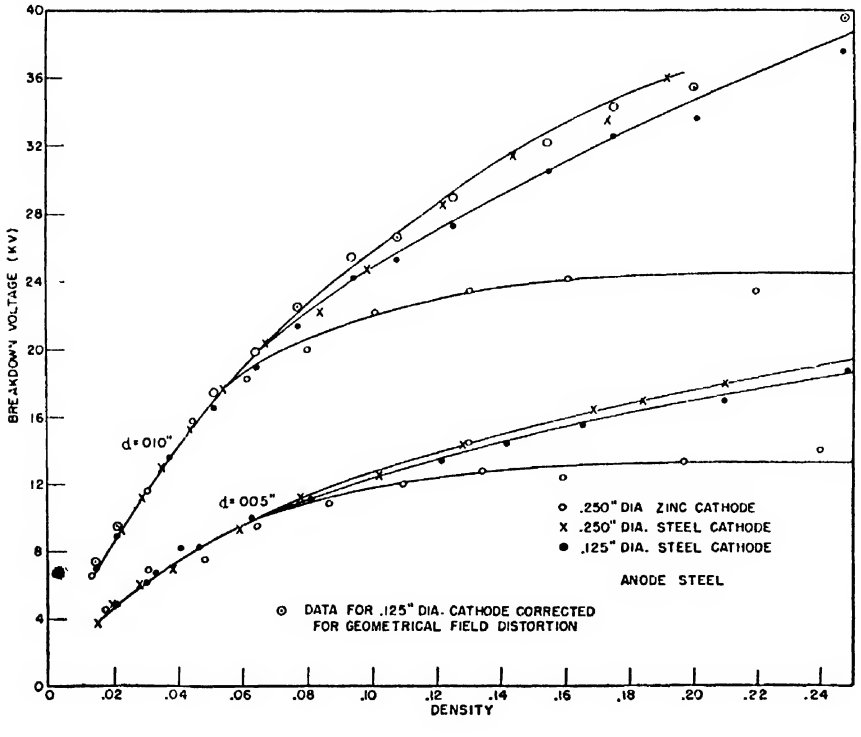


FIG. 12. Breakdown voltage measurements comparing various cathodes.

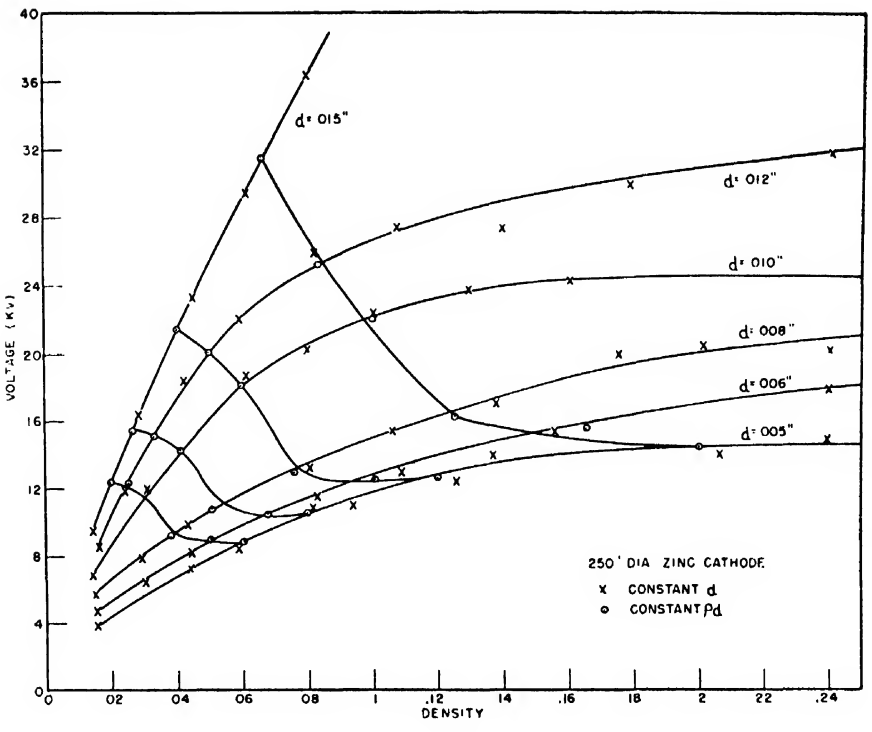


FIG. 13. Breakdown voltage *vs.* density for zinc cathode with gap-length as parameter.

0.125-in. diameter steel electrodes lie lower than those for the 0.250-in. steel electrodes. This is due to the geometrical field distortion for these electrode configurations. The data for the 0.125-in. balls, corrected for geometrical field distortion, yield the same results as for 0.250-in. cathodes (cf. Fig. 12). The breakdown voltage is not influenced by changing the anode metal or by using a 0.125-in. diameter anode instead of the 0.250-in. one.

In Fig. 13 additional breakdown curves for zinc cathodes are given which illustrate the very early failure of Paschen's law.

CURRENT MEASUREMENTS AT HIGH PRESSURES

The prebreakdown currents that flow, when only the α- and γ-type processes of Townsend are present, are proportional to the cathode illumination and are not self-sustaining. If the current is initiated by field emission, there will be a resulting current that does not depend on cathode illumination. A study of prebreakdown currents was initiated, therefore, to check the "field emission hypothesis" suggested above.

A single-stage current amplifier using a Victoreen Instrument Company VX-41A-type tube was constructed to measure the current. The amplifier was operated entirely by self-enclosed batteries with associated meters and adjustable resistors so that changes in battery voltage could be compensated for without change in the calibration of the amplifier (Fig. 14). The indicating instrument used was a 0–10 microammeter. The zero drift was negligible and the calibration remained constant over long periods of time.

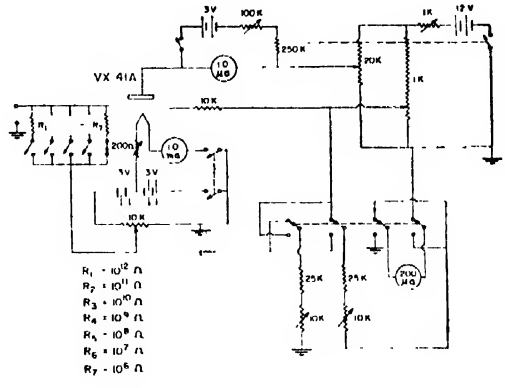


FIG. 14. Current amplifier.

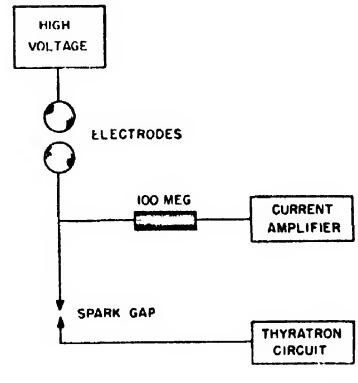


FIG. 15. Protection circuit used in conjunction with current amplifier.

The battery drain was small enough to keep the rate of change of the battery voltage negligible for several hours.

In order to use the protecting circuit described earlier, it was necessary to provide for a triggering voltage in a manner not disturbing to the current measurements (Fig. 15). If breakdown should occur, the small gap fires and supplies the necessary triggering voltage for current limitation and protection of the instrument.

The displacement current charging the electrode capacity of the breakdown gap caused considerable trouble. Instabilities in the high voltage supply, causing displacement current to flow, made the current readings erratic even though the line voltage was regulated by a Sorensen constant voltage transformer. To minimize this difficulty the measurements were made at night. During the course of a particular run (at each

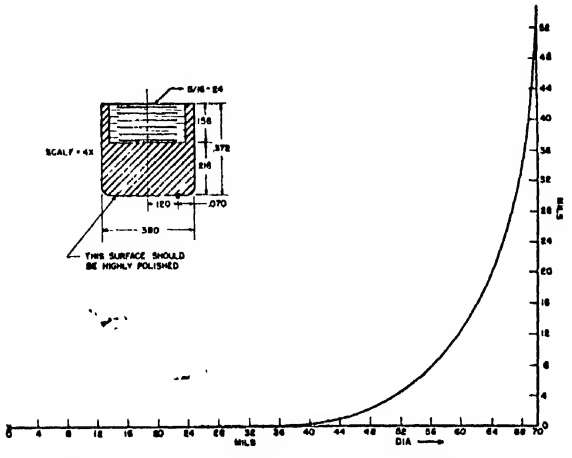


FIG. 16. Electrodes shaped according to Rogowsky's curves (see reference 8).

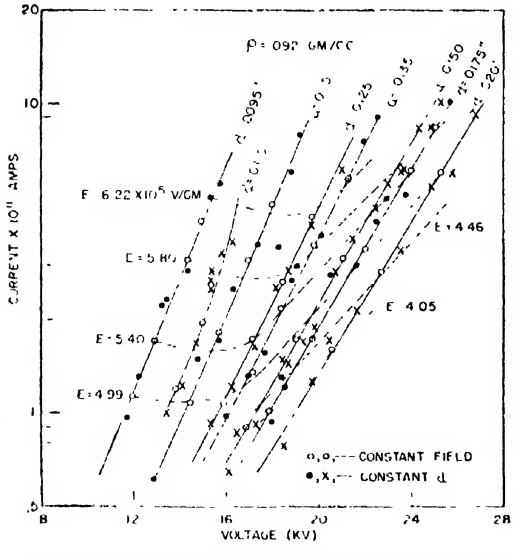


FIG. 17. Prebreakdown current *vs.* applied voltage for density of 0.092 g/cc.

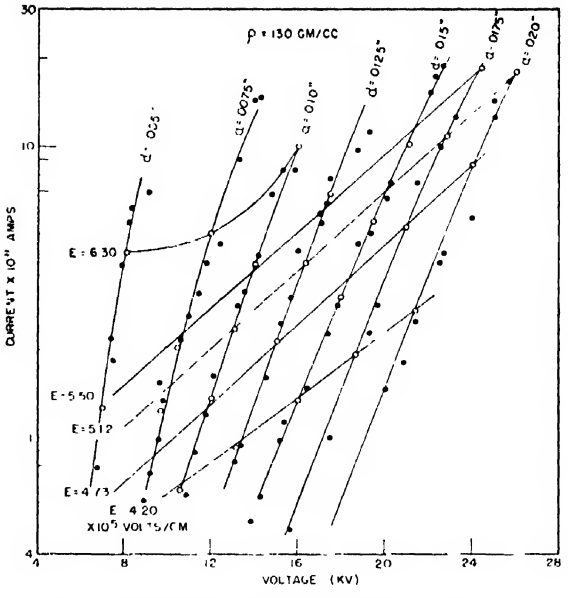


FIG. 18. Prebreakdown current *vs.* applied voltage for density of 0.130 g/cc.

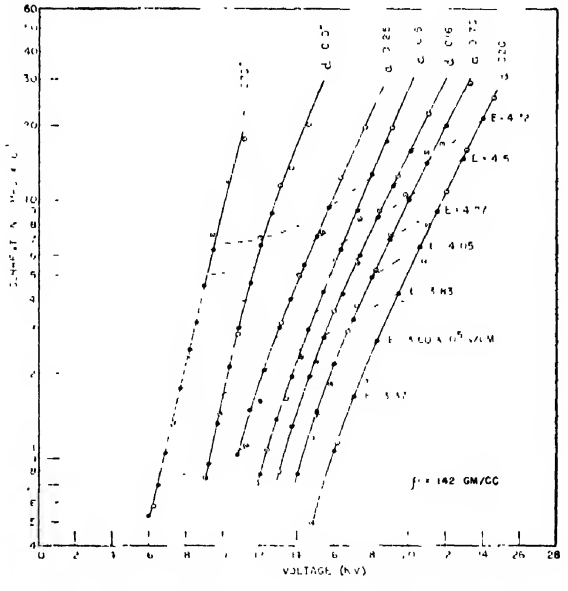


FIG. 19. Prebreakdown current *vs.* applied voltage for density of 0.142 g/cc.

setting of the applied voltage) it was necessary to wait until the electrode capacity charged up to the new voltage before a reading could be taken. These waiting times amounted to several minutes in the case of the smallest currents. A more sentitive indicating instrument would allow the use of smaller grid resistances and reduce this trouble.

The breakdown vessel was directly attached to the amplifier to eliminate disturbances introduced by connecting cables, and provision was made to bring out an insulated connecting wire from the lower electrode

through a pressure seal of the same construction as used at the high voltage terminal (cf. Fig. 1).

Steel electrodes shaped to give a uniform field‡ were used to provide a more definite surface area than that offered by spherical electrodes (Fig. 16).[8] Designed for a maximum spacing of 0.020 in., the electrodes then serve also for smaller spacings.

As other workers have already reported,[9] field emis-

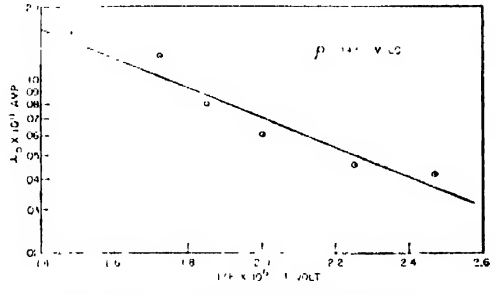


FIG. 20. i_0 *vs.* $1/E$ for density of 0.092 g/cc.

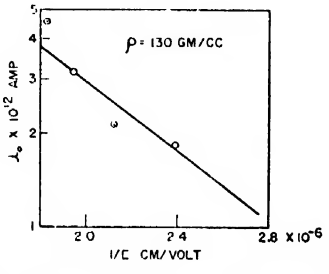


FIG. 21. i_0 *vs.* $1/E$ for density of 0.130 g/cc.

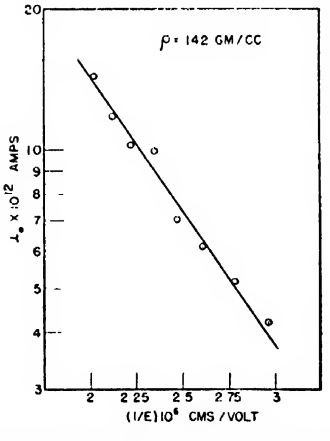


FIG. 22. i_0 *vs.* $1/E$ for density of 0.142 g/cc.

sion currents are very sensitive to the nature of the surface and the electrodes have to be conditioned. The main conditioning procedure consisted in sparking, with the protection circuit preventing excessive burning of the electrodes.

‡ Made and polished by the A. D. Jones Optical Company, Cambridge, Massachusetts.

[8] W. Rogowski, Archiv. f. Elektrotechnik **12**, 1 (1923).

[9] R. A. Millikan and C. F. Eyring, Phys. Rev. **27**, 51 (1926).

TABLE I. Measured values for the field emission coefficients and for α as a function of the applied field.

	ρ = 0.092 g/cc	
	E	α
	4.05×10^5 v/cm	71.5 cm^{-1}
$a=1.21\times10^{-11}$ amp.	4.46	83.9
	4.99	91
$b=1.35\times10^6$ v/cm	5.4	97
	5.8	98.5
	6.22	106.2
	ρ = 0.130 g/cc	
	E	α
$a=4.2\times10^{-11}$ amp.	4.2×10^5 v/cm	54 cm^{-1}
	4.73	74
$b=1.32\times10^6$ v/cm	5.12	77
	5.50	85
	ρ = 0.142 g/cc	
	E	α
	3.37×10^5 v/cm	25.8 cm^{-1}
	3.60	32.2
$a=2.1\times10^{-10}$ amp.	3.83	37.5
	4.05	44.1
	4.27	44.3
$b=1.345\times10^6$ v/cm	4.5	52
	4.72	56.4
	4.95	58.75

Combining the usual equation for field emission with the Townsend equation for impact ionization one obtains:

$$i = ae^{-b/E}e^{\alpha d}. \tag{3}$$

In Figs. 17 to 19 the logarithm of the current is plotted as a function of the voltage for different gap-lengths and curves of constant field strength are drawn. It is noted that straight lines result in this logarithmic plot as anticipated, but only for the longer gap-lengths. It is possible to calculate the Townsend first coefficient

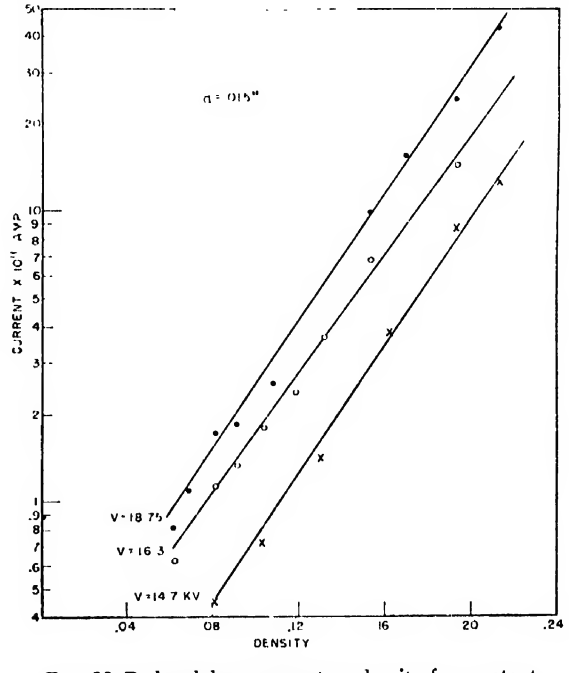


FIG. 23. Prebreakdown current *vs.* density for constant voltage and gap-length.

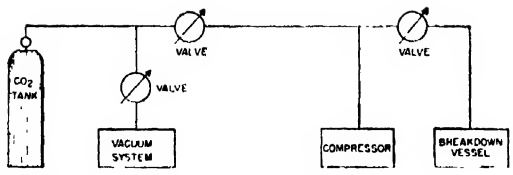


FIG. 24. Experimental apparatus for liquid breakdown.

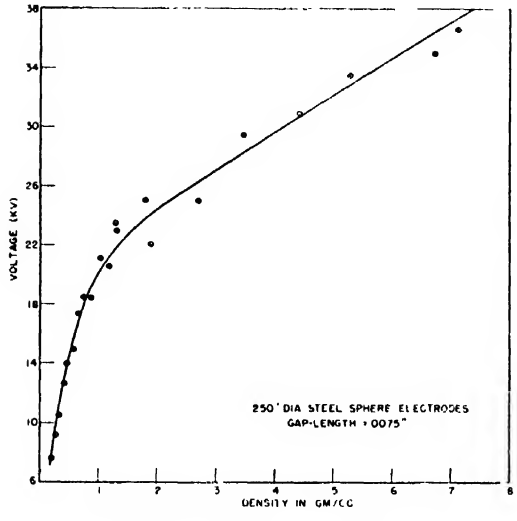


FIG. 25. Breakdown voltage *vs.* density.

α for different strengths from these lines. Then by extrapolation of the lines of constant E to zero voltage (which is the same as zero d) one obtains

$$i_0 = ae^{-b/E} = i_0(E) \quad \text{(cf. Eq. 3)}, \tag{4}$$

and a plot of i_0 as a function of $1/E$ gives the constants a and b (Figs. 20–22). Again a straight line results showing that also the exponential dependence on $1/E$ is fulfilled. Thus the current measurements demonstrate that field emission occurs under these high pressure conditions.

The values of a, b, and α obtained for these gas densities are given in Table I.

If the applied field strength is constant, the current increases with increasing density as shown by Fig. 23. Since α decreases, and b is constant, this is due to an increasing a, as already seen in Table I.

For small gap-spacings the current is larger than predicted by Eq. (3) as can be seen in Figs. 17 to 19, since the curves of constant field strength are no longer straight lines.

BREAKDOWN IN LIQUID CO_2

To make measurements in the liquid phase, a special compressor was constructed to compress the CO_2 gas which is delivered from the tank at saturation pressure. Alternatively, the breakdown vessel was cooled to about 1°C below the tank temperature, in order to condense the CO_2. This latter method is not as flexible as the first one mentioned for establishing a desired pressure above the liquid phase. A diagram of the experimental apparatus is shown in Fig. 24. To study the effect of pressure on the breakdown strength, a thermostat-controlled oil bath was constructed in which the breakdown vessel could be immersed. (No measurements were made in the liquid-vapor equilibrium region where two phases are present.) Measurements were made at the saturation pressure, heating from room temperature and with either gas or liquid CO_2 filling the breakdown vessel while the vessel was sealed off to prevent density change. In both cases there was no change in breakdown voltage with temperature. Therefore, if the density is constant for a given gap-length the breakdown voltage is independent of temperature in the range covered (for gas, 25°–55°C; for liquid, 25°–35°C).

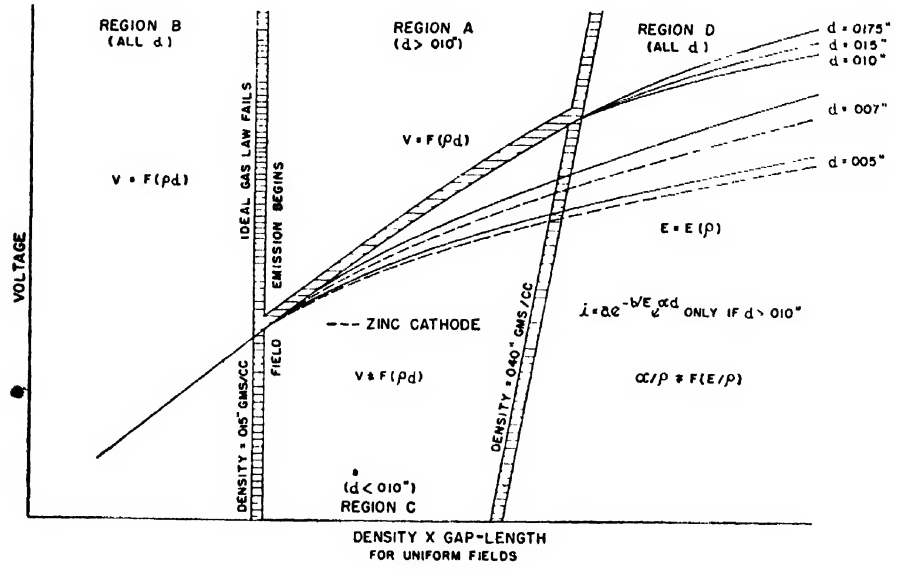


FIG. 26. Breakdown voltage *vs.* ρd (schematic).

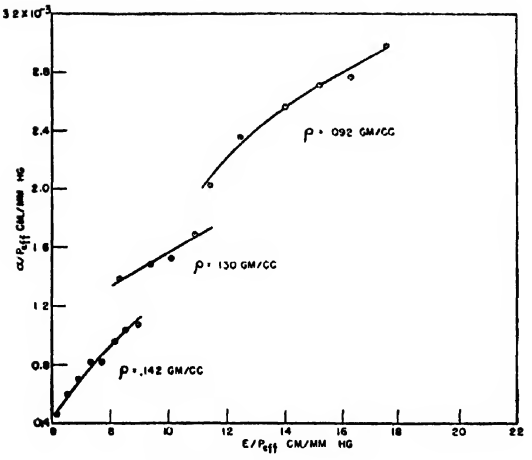


FIG. 27. α/P_{eff} *vs.* E/P_{eff}.

By heating the liquid to 33°C, that is, 2° above the critical temperature, it was possible to change the density continuously from that corresponding to the liquid to that of the vapor. A continuous variation in the breakdown voltage results (Fig. 25).

DISCUSSION

Breakdown Measurements

The most interesting result of these breakdown measurements is the discontinuity in the slope of the ρd curves (Fig. 8) which becomes pronounced above densities $\simeq$0.015 g/cc. This effect corresponds to the branching of the breakdown voltage characteristics plotted in Fig. 26 which occurs at $\rho=0.015$ g/cc. The large departures from Paschen's law for smaller gap-lengths indicate the importance of a new process that aids the breakdown but does not satisfy the similarity law requirements. The onset of this process is partially determined by the cathode material (cf. Fig. 8 with Fig. 13), and its presence is also revealed by the pre-breakdown current which was larger than predicted by Eq. (3) for d's $<$0.010 in. (Figs. 17 to 19).

The explanation of this effect is not obvious. One possibility is that space charge becomes important for small gap-lengths. This would increase the effective ionization coefficient and, in addition, raise the field at the cathode which increases the probability of secondary emission of electrons (γ-process) and the field emission. A solution of the field equation does not predict that an appreciable field distortion occurs for current densities averaged from the measured current; however, field emission does not occur uniformly, and therefore the current densities may be very much higher than assumed.

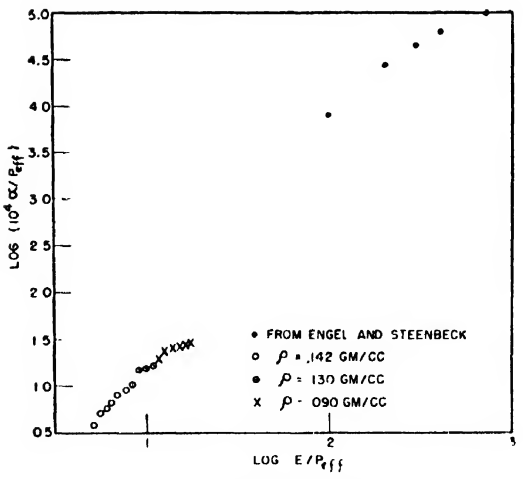


FIG. 28. Present data compared with that given by Engel and Steenbeck (see reference 10) for $\log_{10}(10^4\alpha/P_{eff})$ *vs.* $\log_{10}E/P_{eff}$.

Current Measurements

At low pressures the relation $\alpha/\rho=f(E/\rho)$ holds, but at high pressures, as Fig. 27 indicates, α/ρ is not uniquely determined by E/ρ. This might be expected because the interaction of the molecules will increase the energy loss suffered by the electrons through the excitation of vibrations.

In order to make comparisons with low pressure data, it is convenient to define an effective pressure as

$$P_{eff}=P_0\rho/\rho_0=3.87\times10^5\rho \quad [\text{mm Hg}] \qquad (5)$$

if P_0 and ρ_0 are measured at S.T.P., where the ideal gas law is valid. Low pressure data for α/P *vs.* E/P given by Engel and Steenbeck[10] are combined with the present data in Fig. 28.

The probability of impact ionization before breakdown decreases rapidly with increasing gas density at the highest density studied (Fig. 27).

The increase of the field emission constant a with increasing density may be due to the increasing dielectric constant and its effect on the image force field of the escaping electrons.

The author is grateful to Professor A. von Hippel for suggesting this problem and for constant advice and encouragement during the progress of the work. The many helpful suggestions given by members of the High Voltage Group of this laboratory are also appreciated.

[10] M. Steenbeck and A. v. Engel, *Elektrische Gasentladungen* (Verlag Julius Springer, Berlin, 1932), Vol. 1, p. 106.

A Method for the Numerical Calculation of Hydrodynamic Shocks

J. VonNeumann and R. D. Richtmyer
Institute for Advanced Study, Princeton, New Jersey
(Received September 26, 1949)

The equations of hydrodynamics are modified by the inclusion of additional terms which greatly simplify the procedures needed for stepwise numerical solution of the equations in problems involving shocks. The quantitative influence of these terms can be made as small as one wishes by choice of a sufficiently fine mesh for the numerical integrations. A set of difference equations suitable for the numerical work is given, and the condition that must be satisfied to insure their stability is derived

I. INTRODUCTION

In the investigation of phenomena arising in the flow of a compressible fluid, it is frequently desirable to solve the equations of fluid motion by stepwise numerical procedures, but the work is usually severely complicated by the presence of shocks. The shocks manifest themselves mathematically as surfaces on which density, fluid velocity, temperature, entropy and the like have discontinuities; and clearly the partial differential equations governing the motion require boundary conditions connecting the values of these quantities on the two sides of each such surface. The necessary boundary conditions are, of course, supplied by the Rankine-Hugoniot equations, but their application is complicated because the shock surfaces are in motion relative to the network of points in space-time used for the numerical work, and the differential equations and boundary conditions are non-linear. Furthermore, the motion of the surfaces is not known in advance but is governed by the differential equations and boundary conditions themselves. In consequence, the treatment of shocks requires lengthy computations (usually by trial and error) at each step, in time, of the calculation.

We describe here a method for automatic treatment of shocks which avoids the necessity for application of any such boundary conditions. The approximations in it can be rendered as accurate as one wishes, by suitable choice of interval sizes and other parameters occurring in the method. It treats all shocks, correctly and automatically, whenever and wherever they may arise.

The method utilizes the well-known effect on shocks of dissipative mechanisms, such as viscosity and heat conduction.[1] When viscosity is taken into account, for example, the shocks are seen to be smeared out, so that the mathematical surfaces of discontinuity are replaced by thin layers in which pressure, density, temperature, etc. vary rapidly but continuously. Our idea is to introduce (artificial) dissipative terms into the equations so as to give the shocks a thickness comparable to (but preferably somewhat larger than) the spacing of the points of the network. Then the differential equations (more accurately, the corresponding difference equations) may be used for the entire calculation, just as though there were no shocks at all. In the numerical results obtained, the shocks are immediately evident as near-discontinuities that move through the fluid with very nearly the correct speed and across which pressure, temperature, etc. have very nearly the correct jumps.

It will be seen that for the assumed form of dissipation (and, indeed, for many others as well), the Rankine-Hugoniot equations are satisfied, provided the thickness of the shock layers is small in comparison with other physically relevant dimensions of the system. We then consider one way in which the transition from differential to finite-difference equations can be made and we discuss the mathematical stability of these equations. It will be seen that the dissipative terms have the effect of making the stability condition more stringent than the familiar one of Courant, Friedrichs, and Lewy,[2] but not seriously so if the amount of dissipation introduced is only enough to produce a shock thickness comparable with the spatial interval length of the network used.

The method has been applied, so far, only to one-dimensional flows, but appears to be equally suited to the study of more complicated flows; where, indeed, shock calculations by direct application of the Hugoniot equations would ordinarily be prohibitively difficult, even for rapid, automatic computers.

The discussions which follow are primarily intended to give a complete picture of the ideas and mathematical procedures involved. In some places (Chapter VII, also the essential inferences from some of the material of Chapters IV, V) the mathematical discussions are, however, carried through only with a view to give a complete chain of the necessary procedure, but not with all the detail that rigorous proofs in a primarily mathematical paper would require. The reason for doing this was partly desire to avoid inconvenient length, partly that of not wishing to have to select now the precise degree of generality for the

[1] Lord Rayleigh (Proc. Roy. Soc. **A84**, 247 (1910)) and G. I. Taylor (Proc. Roy. Soc. **A84**, 371 (1910)) showed, on the basis of general thermodynamical considerations, that dissipation is necessarily present in shock waves. Later, R. Becker (Zeits. f. Physik **8**, 321 (1922)) gave a detailed discussion of the effects of heat conduction and viscosity. Recently, L. H. Thomas (J. Chem. Phys. **12**, 449 (1944)) has investigated these effects further in terms of the kinetic theory of gases.

[2] Courant, Friedrichs, and Lewy, Math. Ann. **100**, 32 (1928). It is in this important paper that these authors first published their discovery of the conditional stability of the difference-equation integration method for partial differential equations.

validity of the method. It seems preferable to reserve such discussions for later occasions.

The validity of our methods has been tested empirically on various computational applications. These are partly still under analysis, and will be published and discussed elsewhere.

II. THE BASIC EQUATIONS

Consider a one-dimensional fluid motion. Let x be the Lagrangean coordinate, and $X=X(x, t)$ be the Eulerian coordinate. That is, $X(x, t)$ gives the position, at time t, of a fluid element that was initially at position x. Let $\rho_0(x)$ be the initial density, so that V and U, given by

$$V(x, t)=(1/\rho_0(x))(\partial X/\partial x) \tag{1}$$

and

$$U(x, t)=\partial X/\partial t, \tag{2}$$

are the specific volume and fluid velocity, respectively. The equations of motion, of energy, and of continuity are:

$$\rho_0(\partial U/\partial t)=-(\partial/\partial x)(p+q), \tag{3}$$

$$(\partial \mathcal{E}/\partial t)+(p+q)(\partial V/\partial t)=0, \tag{4}$$

and

$$\rho_0(\partial V/\partial t)=(\partial U/\partial x). \tag{5}$$

In these equations, $p=p(x, t)$ is the ordinary (or static) fluid pressure and $\mathcal{E}=\mathcal{E}(x, t)$ is the internal energy per unit mass. A connection between $\mathcal{E}$, p, V is established by an equation of state, which will be assumed, for the purpose of illustration, to have the form

$$\mathcal{E}=(pV)/(\gamma-1) \tag{6}$$

which holds, for example, in the case of a perfect gas. γ is a constant >1. It is supposed that the dissipative mechanism can be represented by the additional term q in the pressure, which is assumed to be negligibly small, except in the neighborhood of the shock.

III. THE EXPRESSION FOR q

The dissipation is introduced for purely mathematical reasons. Therefore q may be taken as any convenient function of p, V, etc. and their derivatives, provided that the following requirements are met:

1. The Eqs. (3), (4), and (5) must possess solutions without discontinuities.
2. The thickness of the shock layers must be everywhere of the same order as the interval length Δx used in the numerical computation, independently of the strength of the shock and of the condition of the material into which it is running.
3. The effect of the terms containing q in (3) and (4) must be negligible outside of the shock layers.
4. The Hugoniot equations must hold when all other dimensions characterizing the flow are large compared to the shock thickness.

We shall show that the expression

$$q=-\frac{(\rho_0 c\Delta x)^2}{V}\frac{\partial V}{\partial t}\left|\frac{\partial V}{\partial t}\right| \tag{7}$$

meets the requirements. c is a dimensionless constant near unity. The dissipative mechanism is essentially a non-linear viscosity as can be seen more clearly if (7) is written equivalently (for the one-dimensional case) as

$$q=-\frac{(c\Delta x)^2}{V}\frac{\partial U}{\partial x}\left|\frac{\partial U}{\partial x}\right| \tag{8}$$

by use of (5). To show that the expression (7) meets the stated requirements, we consider a steady-state shock.

IV. STEADY-STATE PLANE SHOCK

Imagine a long pipe containing a fluid initially in equilibrium (thermally and mechanically), into which a piston is pushing from one end with constant speed, as shown in Fig. 1. In the absence of dissipation the specific volume, V, and the fluid velocity, U, are, at a given instant, as shown by the solid curves in the graphs, whereas in the presence of dissipation they are as shown by the dashed curves. In either case, the shock is steady, at least approximately, after it has gone to a sufficiently great distance from the initiating piston. Then U, V, $\mathcal{E}$, etc. depend on x and t only through the combination

$$w=x-st, \tag{9}$$

where s is the speed of the shock relative to the original, or Lagrangean, coordinates. We suppose that ρ_0 and Δx are constants (independent of x).

It is convenient to define

$$M=\rho_0 s \tag{10}$$

—in a co-moving coordinate system, M is the mass crossing unit area in unit time—whereupon Eqs. (3)–(5) become:

$$M(dU/dw)=(d/dw)(p+q), \tag{11}$$

$$(d\mathcal{E}/dw)+(p+q)(dV/dw)=0, \tag{12}$$

and

$$-M(dV/dw)=dU/dw. \tag{13}$$

Then, (11) and (13) give:

$$-M^2(dV/dw)=(d/dw)(p+q); \tag{14}$$

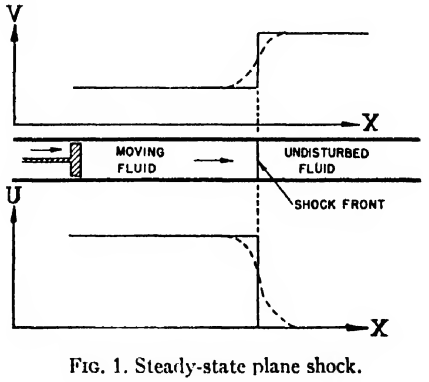


FIG. 1. Steady-state plane shock.

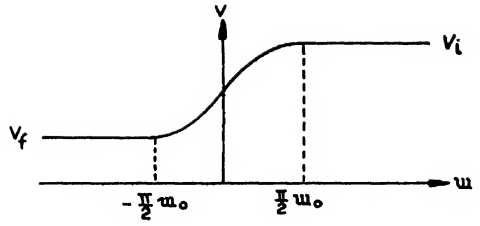


FIG. 2. Specific volume in steady shock.

and (12) and (14) give

$$\frac{d\mathcal{E}}{dw}+\frac{d}{dw}[(p+q)V]+M^2V\frac{dV}{dw}=0. \tag{15}$$

The solutions of (13), (14), (15) are:

$$MV+U=C_1 \tag{16}$$

$$M^2V+p+q=C_2 \tag{17}$$

$$\mathcal{E}+(p+q)V+\tfrac{1}{2}M^2V^2=C_3, \tag{18}$$

where C_1, C_2, and C_3 are constants.

Let the initial and final values be denoted by:

$$\text{As } w\to\infty;\quad V\to V_i,\quad p\to p_i,\quad \mathcal{E}\to\mathcal{E}_i,\quad q\to 0; \tag{19}$$

$$\text{As } w\to-\infty;\quad V\to V_f,\quad p\to p_f,\quad \mathcal{E}\to\mathcal{E}_f,\quad q\to 0. \tag{20}$$

Then (17) gives:

$$M^2(V_i-V_f)=P_f-P_i; \tag{21}$$

and (18) and (21) together give:

$$\mathcal{E}_f-\mathcal{E}_i=\tfrac{1}{2}(P_i+P_f)(V_i-V_f). \tag{22}$$

(21) and (22) are the equations of Hugoniot and are seen to be independent of the amount and form of the dissipation, provided that $q\to 0$ as $w\to\pm\infty$. The physical reason for this is that the Hugoniot equations are direct consequences of the conservation laws of mass, momentum and energy, and the form of dissipation assumed is such as top reserve the over-all conservation of these quantities. These laws require that in a shock a certain amount of mechanical energy be converted irreversibly into heat. In the steady state, the motion adjusts itself, in the shock layer, until precisely that amount of work is done against the pressure q according to Eq. (3) and is converted into heat according to Eq. (4).

To investigate the shape of the shock, we first look for solutions satisfying

$$(\partial V/\partial t)\leq 0 \quad \text{or} \quad (dV/dw)\geq 0. \tag{23}$$

This is normally the case for a shock moving to the right. Then (7) can be written:

$$qV=+(Mc\Delta x)^2(dV/dw)^2. \tag{24}$$

From (17) and (18)

$$\mathcal{E}-\tfrac{1}{2}M^2V^2=C_3-C_2V, \tag{25}$$

so that by (6),

$$pV=[(\gamma-1)/2]M^2V^2+C_4V+C_5, \tag{26}$$

whereupon (17) gives:

$$qV=C_2-[(\gamma+1)/2]M^2V^2-C_4V-C_5. \tag{27}$$

The right member of (27) is a quadratic in V that vanishes for $V=V_i$ and for $V=V_f$, so that, clearly,

$$(Mc\Delta x)^2\left(\frac{dV}{dw}\right)^2=qV=\frac{\gamma+1}{2}M^2(V_i-V)(V-V_f). \tag{28}$$

(Equation (28) can also be obtained from (27) by direct use of the Hugoniot equations and the equations that fix C_2, C_4, C_5.)

To solve (28), put:

$$\psi=V-\frac{V_i+V_f}{2},\quad \psi_0=\frac{V_i-V_f}{2},\quad \varphi=\frac{\psi}{\psi_0}, \tag{29}$$

so that

$$c\Delta x\frac{d\psi}{dw}=[(\gamma+1)/2]^{\frac{1}{2}}(\psi_0^2-\psi^2)^{\frac{1}{2}}, \tag{30}$$

or

$$c\Delta x\frac{d\varphi}{dw}=[(\gamma+1)/2]^{\frac{1}{2}}(1-\varphi^2)^{\frac{1}{2}}. \tag{31}$$

Therefore,

$$w=[2/(\gamma+1)]^{\frac{1}{2}}c\Delta x\int\frac{d\varphi}{(1-\varphi^2)^{\frac{1}{2}}}=w_0\arcsin\varphi \tag{32}$$

where

$$w_0=[2/(\gamma+1)]^{\frac{1}{2}}c\Delta x. \tag{33}$$

Finally,

$$\psi=V-\frac{V_i+V_f}{2}=\frac{V_i-V_f}{2}\sin\frac{w}{w_0}. \tag{34}$$

Because of our initial assumption (23) that $dV/dw\geq 0$, we can use only a half wave of the solution (34), but this half wave can be pieced together with two other particular solutions,

$$V\equiv V_i \quad \text{and} \quad V\equiv V_f \tag{35}$$

(they satisfy Eq. (28)), to make the composite continuous solution depicted in Fig. 2. w_0 is a measure of the thickness of the shock, and is of order Δx, provided c is of order unity, independently of the strength of the shock and conditions ahead of it, by Eq. (33). Throughout most of the system q is negligible in comparison with the ordinary pressure, p, because of the factor $(\Delta x)^2$ in (7); but in the shock layer q becomes comparable with p because of the abnormally large value of $\partial V/\partial t$ there.

Expression (7) thus meets all requirements.

It may be noted that if we had looked instead for solutions having $dV/dt>0$, no solution would have been found for which all quantities are continuous and bounded, because in that case the opposite sign would occur in (28), leading to the hyperbolic instead of ordi-

nary sine function. Therefore negative shocks do not exist in the steady state, for our equations. Negative shocks do not exist in the physical world, either, so that our expression (7) is satisfactory from this point of view.

V. STABILITY OF THE DIFFERENTIAL EQUATIONS

Suppose that at some instant there is superposed on a desired solution $U(x, t)$, $V(x, t)$, etc., a small perturbation δU, δV, etc. It is of interest to know whether the perturbation grows with increasing time. To investigate this we replace U, V, etc. by $U+\delta U$, $V+\delta V$, etc. in (3), (4), (5), (8), after first rewriting (4), by use of (6), in the form:

$$[\gamma p+(\gamma-1)q](\partial V/\partial t)+V(\partial p/\partial t)=0. \tag{36}$$

We obtain in this way the equations of first variation:

$$\rho_0(\partial/\partial t)\delta U=-(\partial/\partial x)(\delta p+\delta q), \tag{37}$$

$$\left.\begin{aligned}\frac{\partial V}{\partial t}[\gamma\delta p+(\gamma-1)\delta q]+[\gamma p+(\gamma-1)q]\frac{\partial}{\partial t}\delta V\\ +V\frac{\partial}{\partial t}\delta p+\frac{\partial p}{\partial t}\delta V=0,\end{aligned}\right\} \tag{38}$$

$$\delta q=\frac{(c\Delta x)^2}{V^2}\frac{\partial U}{\partial x}\cdot\left|\frac{\partial U}{\partial x}\right|\delta V-2\frac{(c\Delta x)^2}{V}\cdot\left|\frac{\partial U}{\partial x}\right|\frac{\partial}{\partial x}\delta U, \tag{39}$$

$$\rho_0\frac{\partial}{\partial t}\delta V=-\frac{\partial}{\partial x}\delta U. \tag{40}$$

In writing the last term of (39) we have assumed that the perturbation is not large enough to alter the sign of $\partial U/\partial x$. (37) to (40) are a set of simultaneous linear differential equations for δU, δV, δp, δq. Their coefficients depend on the desired solution U, V, p, q, and are thought of as smoothly varying functions of x, t. We shall be concerned with rapidly varying perturbations. We therefore treat the coefficients of (37) to (40) as constants in a small region and look for solutions having the form:

$$\delta U=\delta U_0e^{ikx+\alpha t},\quad \delta V=\delta V_0e^{ikx+\alpha t},\text{ etc.,} \tag{41}$$

where δU_0, δV_0, δp_0, δq_0, k, and α are constant and k is real. Substitution of (41) into (37) to (40) leads to four simultaneous homogeneous linear equations in δU_0, δV_0, δp_0, δq_0. The vanishing of the determinant of these equations establishes an equation connecting α and k. By solving this equation, for given k, and examining the real part of α, we can determine whether a given Fourier component (41) of the perturbation grows with increasing time. This program is readily carried out, but we omit the details in the interest of brevity. The determinantal equation is:

$$\begin{aligned}(\alpha\rho_0)^2\gamma\frac{\partial V}{\partial t}+2\frac{\rho_0\alpha}{V}\frac{\partial V}{\partial t}(kc\Delta x)^2\left|\frac{\partial U}{\partial x}\right|\\ -\frac{1}{V^2}\frac{\partial V}{\partial t}(kc\Delta x)^2\frac{\partial U}{\partial x}\cdot\left|\frac{\partial U}{\partial x}\right|+k^2\alpha[\gamma p+(\gamma-1)q]\\ +\rho_0^2\alpha^3V+2\rho_0\alpha^2(kc\Delta x)^2\left|\frac{\partial U}{\partial x}\right|\\ -\frac{\alpha}{V}(kc\Delta x)^2\frac{\partial U}{\partial x}\cdot\left|\frac{\partial U}{\partial x}\right|+k^2\frac{\partial p}{\partial t}=0.\end{aligned} \tag{42}$$

If we restrict our attention to Fourier components with very large k, only certain terms of (42) need be retained, the others being negligible either by viture of being of lower order in k or by virtue of being of lower order in α. Two cases are distinguished: in shock regions we retain all terms in (42); in normal regions we drop the dissipative terms (those containing Δx). In the two regions, the dominant terms, in the sense explained, give: shock regions:

$$\alpha=-2\frac{(kc\Delta x)^2}{\rho_0V}\left|\frac{\partial U}{\partial x}\right| \tag{43}$$

normal regions:

$$\alpha^2=-\frac{k^2}{\rho_0V}\frac{\gamma p}{\rho_0}. \tag{44}$$

It is seen that small disturbances are damped out in the shock layers but propagate without either growth or decay in normal regions. This corresponds to physical reality, so our expression (7) is satisfactory also as regards stability of the resulting differential equations.

We can furthermore identify the terms in the equations of variation that lead to the dominant terms in (42). They give:

shock regions:

$$\frac{\partial}{\partial t}\delta U\approx\sigma\frac{\partial^2}{\partial x^2}\delta U, \tag{45}$$

where

$$\sigma=\frac{2(c\Delta x)^2}{V\rho_0}\left|\frac{\partial U}{\partial x}\right|;$$

normal regions:

$$\frac{\partial^2}{\partial t^2}\delta U\approx s_0^2\frac{\partial^2}{\partial x^2}\delta U, \tag{46}$$

where

$$s_0^2=\frac{\gamma p}{\rho_0^2V}.$$

Thus our system has the character of a diffusion equation in the shock layers and of a sound equation elsewhere.

VI. FINITE DIFFERENCE EQUATIONS

There are many systems of finite difference equations equivalent to the differential equations, but we shall restrict the discussion to one of the simplest. Certain other systems are much superior from the point of view of stability but are more complicated from the point of view of numerical solution. Systems of the latter type will be discussed elsewhere.

Let the points of a rectangular network with spacings Δx and Δt be denoted by x_l, $t^n (l=0, 1, 2, \cdots, L; n=0, 1, 2, \cdots)$. We shall also have occasion to deal with intermediate points, having coordinates $x_{l+\frac{1}{2}} \overset{\text{def.}}{=} \frac{1}{2}(x_{l+1}+x_l)$ etc. To facilitate the writing, we introduce abbreviations such as:

$$V_{l+\frac{1}{2}}^{n} = V(x_{l+\frac{1}{2}}, t^n) \text{ etc.} \tag{47}$$

The difference equations corresponding to (3), (5), (8)' and (36) are:

$$\rho_0 \frac{U_l^{n+\frac{1}{2}} - U_l^{n-\frac{1}{2}}}{\Delta t} = \frac{p_{l+\frac{1}{2}}^{n} + q_{l+\frac{1}{2}}^{n-\frac{1}{2}} - p_{l-\frac{1}{2}}^{n} - q_{l-\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta x}, \tag{48}$$

$$\rho_0 \frac{V_{l+\frac{1}{2}}^{n+1} - V_{l+\frac{1}{2}}^{n}}{\Delta t} = \frac{U_{l+1}^{n+\frac{1}{2}} - U_l^{n+\frac{1}{2}}}{\Delta x}, \tag{49}$$

$$q_{l+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{2(c\Delta x)^2}{V_{l+\frac{1}{2}}^{n} + V_{l+\frac{1}{2}}^{n+1}} \cdot \frac{(U_{l+1}^{n+\frac{1}{2}} - U_l^{n+\frac{1}{2}}) \cdot |U_{l+1}^{n+\frac{1}{2}} - U_l^{n+\frac{1}{2}}|}{(\Delta x)^2}, \tag{50}$$

and

$$\left[\gamma \frac{p_{l+\frac{1}{2}}^{n+1} + p_{l+\frac{1}{2}}^{n}}{2} + (\gamma-1) q_{l+\frac{1}{2}}^{n+\frac{1}{2}}\right] \frac{V_{l+\frac{1}{2}}^{n+1} - V_{l+\frac{1}{2}}^{n}}{\Delta t} + \frac{V_{l+\frac{1}{2}}^{n+1} + V_{l+\frac{1}{2}}^{n}}{2} \frac{p_{l+\frac{1}{2}}^{n+1} - p_{l+\frac{1}{2}}^{n}}{\Delta t} = 0. \tag{51}$$

These equations are correct to second order of small quantities Δx and Δt, except for the terms containing q in (48) which are only correct to the first order; but these terms are negligible except in the shock layers and are physically artificial in any case.

For numerical solution, suppose that all quantities are known for superscript n or less. Compute $U_l^{n+\frac{1}{2}}$ from (48) for each l; compute $V_{l+\frac{1}{2}}^{n+1}$ from (49) for each l; compute $q_{l+\frac{1}{2}}^{n+\frac{1}{2}}$ from (50) for each l; compute $p_{l+\frac{1}{2}}^{n+\frac{1}{2}}$ from (51) for each l; this completes a cycle. Boundary conditions are needed, an example being (rigid walls at ends of a tube): $U_0^{n+\frac{1}{2}}=0$, $U_L^{n+\frac{1}{2}}=0$.

VII. STABILITY OF THE DIFFERENCE EQUATIONS

Equations (48) to (51), being only approximations to the differential equations, cannot be expected to give all features of the solution with precision. If U, V are thought of as being expanded in Fourier series with coefficients depending on t, the long-wave-length components are accurately given by (48) to (51), provided Δx and Δt are sufficiently small, but the components whose wave-lengths are of order Δx are always falsified somewhat. This falsification is harmless, provided that all physically relevant components are treated accurately; this requires not only that Δx and Δt be small but also that the physically insignificant components with wave-lengths of order Δx remain small during the entire calculation. It may happen not only that the short-wave-length components are falsified but also that their amplitudes increase with increasing n, in spite of the stability of the *differential* equations. This increase is generally exponential and, if it occurs, quickly makes gibberish of the entire calculation.

The avoidance of such catastrophes, when partial differential equations are approximated by difference equations, has been the subject of study by various investigators, beginning with the fundamental paper of Courant, Friedrichs, and Lewy referred to in reference 1. We shall give a somewhat heuristic discussion of the stability questions met in the present problem.

We again suppose a small perturbation δU, δV, etc., superposed on a smooth solution, and consider the equations of variation of (48) to (51). According to the analysis of Part V, the dominant terms are expected to be (see Eqs. (45) and (46)):

shock regions:

$$\frac{\delta U_l^{n+\frac{1}{2}} - \delta U_l^{n-\frac{1}{2}}}{\Delta t} \approx \sigma \frac{\delta U_{l+1}^{n-\frac{1}{2}} - 2\delta U_l^{n-\frac{1}{2}} + \delta U_{l-1}^{n-\frac{1}{2}}}{(\Delta x)^2}; \tag{52}$$

normal regions:

$$\frac{\delta U_l^{n+\frac{1}{2}} - 2\delta U_l^{n-\frac{1}{2}} + \delta U_l^{n-\frac{3}{2}}}{(\Delta t)^2} \approx s_0^2 \frac{\delta U_{l+1}^{n-\frac{1}{2}} - 2\delta U_l^{n-\frac{1}{2}} + \delta U_{l-1}^{n-\frac{1}{2}}}{(\Delta x)^2}; \tag{53}$$

and this can be verified by writing out the difference equations of variation in detail.

As before, we consider perturbations of the form:

$$\delta U = \delta U_0 e^{ikx+\alpha t}, \text{ etc.}, \tag{54}$$

so that:

$$\delta U_l^{n+\frac{1}{2}} = \delta U_0 \zeta^l \xi^{n+\frac{1}{2}}, \tag{55}$$

where:

$$\zeta = e^{ik\Delta x}, \quad \xi = e^{\alpha \Delta t}. \tag{56}$$

For stability we require that $|\xi| \leq 1$ for all real k.

We consider normal regions first. Substitution of (55) into (53) and cancellation of common factors gives:

$$\xi - 2 + \frac{1}{\xi} = 2\left(\frac{s_0 \Delta t}{\Delta x}\right)^2 (\cos k\Delta x - 1). \tag{57}$$

If we call

$$L=\frac{s_0\Delta t}{\Delta x}, \tag{58}$$

the solution of (57) is

$$\xi_{1,2}=b\pm(b^2-1)^{\frac{1}{2}} \tag{59}$$

where

$$b=1-L^2(1-\cos k\Delta x). \tag{60}$$

According to (60), b is always <1, so there are two cases:

$$-1<b<1, \quad |\xi_1|=|\xi_2|=1; \quad \text{stability}; \tag{61}$$

$$b<-1, \quad |\xi_1|<1<|\xi_2|; \quad \text{instability}. \tag{62}$$

From Eq. (60) it is seen that (61) will hold for all k if and only if

$$L\leq 1. \tag{63}$$

This is the familiar condition for stability of hydrodynamical equations of the form (53).

A similar treatment of Eq. (52) for the shock region yields directly:

$$\xi-1=2\frac{\sigma\Delta t}{(\Delta x)^2}(\cos k\Delta x-1), \tag{64}$$

and the stability condition is clearly that

$$\frac{2\sigma\Delta t}{(\Delta x)^2}\leq 1 \quad \text{or} \quad \Delta t\leq\frac{(\Delta x)^2}{2\sigma}. \tag{65}$$

To interpret this result, we calculate σ according to (45) for the steady shock discussed in Part IV. From (33), (34);

$$\left|\frac{\partial U}{\partial x}\right|=\rho_0\left|\frac{\partial V}{\partial t}\right|=\rho_0 s\left|\frac{\partial V}{\partial w}\right|$$

$$=\rho_0 s\frac{V_s-V_f}{2c\Delta x}((\gamma+1)/2)^{\frac{1}{2}}\cos\frac{w}{w_0}. \tag{66}$$

Therefore

$$\sigma=sc\Delta x\frac{V_s-V_f}{V}((\gamma+1)/2)^{\frac{1}{2}}\cos\frac{w}{w_0}, \tag{67}$$

and the stability condition (65) takes the form:

$$\Delta t\leq\frac{V\Delta x}{2sc(V_s-V_f)}\left(\frac{2}{\gamma+1}\right)^{\frac{1}{2}}\frac{1}{\cos(w/w_0)}, \tag{68}$$

or,

$$\Delta t\leq\frac{\Delta x}{4sc}\left(\frac{2}{\gamma+1}\right)^{\frac{1}{2}}\frac{(\eta+1)/(\eta-1)+\sin(w/w_0)}{\cos(w/w_0)}, \tag{69}$$

by further use of (34), where

$$\eta=V_i/V_f. \tag{70}$$

The quantity η is a measure of the shock strength. Equation (69) shows that different parts of the shock layer (i.e., different values of w) impose different restrictions on Δt. The effective restriction is obtained by replacing the right member of (69) by its minimum value for $-(\pi/2)w_0\leq w\leq(\pi/2)w_0$. The minimum of the last factor in (69) is found to be $2(\eta)^{\frac{1}{2}}/(\eta-1)$, and the stability condition is

$$\Delta t\leq\frac{\Delta x}{2sc}\left(\frac{2}{\gamma+1}\right)^{\frac{1}{2}}\frac{\eta^{\frac{1}{2}}}{\eta-1}. \tag{71}$$

For practical application, it is convenient to express (71) in terms of the quantity L appearing in the normal hydrodynamic stability condition (63). By elimination of p_i from the Hugoniot Eqs. (21), (22), and (8), the speed of the shock is found to be

$$\left.\begin{aligned}s=\frac{M}{\rho_0}=\frac{1}{\rho_0}\left(\frac{2}{(\gamma+1)\eta-(\gamma-1)}\right)^{\frac{1}{2}}\left(\gamma\frac{p_f}{V_f}\right)^{\frac{1}{2}}\\=\left(\frac{2}{(\gamma+1)\eta-(\gamma-1)}\right)^{\frac{1}{2}}s_{0f},\end{aligned}\right\} \tag{72}$$

where s_{0f} is the speed of sound, relative to Lagrangean coordinates, in the material behind the shock. By combining (71) and (72) and use of (58), the stability condition becomes:

$$L_f=\frac{s_{0f}\Delta t}{\Delta x}\leq\frac{1}{2c}\frac{\{[\eta-(\gamma-1)/(\gamma+1)]\eta\}^{\frac{1}{2}}}{\eta-1}. \tag{73}$$

Lastly, the shock strength, and hence η, is generally unknown until the calculation has been performed; it is therefore advisable to replace the right member of (73) by its minimum with respect to η. η can vary in the range $1\leq\eta\leq(\gamma+1)/(\gamma-1)$ (the latter value corresponding to an infinitely strong shock) and the minimum of (73) is attained at the upper end of this range, where the last factor in (73) has the value $\gamma^{\frac{1}{2}}$. Our final, sufficient condition for stability reads:

$$L_f\leq\gamma^{\frac{1}{2}}/2c. \tag{74}$$

This condition has been found to insure stability in test calculations, whereas a serious violation of it leads to trouble. The choice $c=1$ has been found to yield good results in practice for the representation of shocks, in which case the stability condition is at worst slightly more severe than the one that must be observed, anyway (compare Eq. (63)), to insure stability of the motion behind the shock.

Some Non-Linear Systems Permitting Simple Harmonic Motion

JOHN E. BROCK
Department of Mechanics, Washington University, St. Louis, Missouri
(Received October 24, 1949)

By elementary and obvious methods a differential equation is obtained, which in general is non-linear and usually affords a solution which is a simple sinusoidal function of the independent variable (i.e., simple harmonic motion). Two particular examples are studied briefly although no physical realizations are presented. One is led to speculate that new types of oscillators having a sinusoidal output are described by these equations, and that practical applications, such as to frequency metering, may eventually be evolved. However, the writer's speculations along these lines have as yet borne no fruit and his primary object in publishing this material is to bring to the attention of those interested in non-linear phenomena, about which our understanding is so essentially imperfect, an analysis which, though itself possibly trivial, conceivably may prove of service in setting someone else off on a productive investigation.

THE ordinary linear oscillator described by the equation

$$\ddot{x}+\omega^2 x=0, \tag{1}$$

in which $\omega^2>0$, permits simple harmonic motion given by

$$x=A\sin(\omega t+\phi), \tag{2}$$

in which A and ϕ are arbitrary constants, determined by the initial conditions of the motion and independent of ω^2.

This is a special case of the more general oscillator described by

$$F[-(\ddot{x}/x),\quad x^2-(x\dot{x}^2/\dot{x})]=0, \tag{3}$$

in which F represents an arbitrary function of two variables. Equation (3) possesses the solution given by Eq. (2) subject to the condition

$$F(\omega^2,A^2)=0. \tag{4}$$

That is, frequency ω and amplitude A are related as indicated by Eq. (4). Equation (3) may also possess solutions other than that given by Eq. (2).

Equation (3) is derived as follows. From Eq. (2) we also have

$$\dot{x}=A\omega\cos(\omega t+\phi), \tag{5}$$

$$\ddot{x}=-A\omega^2\sin(\omega t+\phi). \tag{6}$$

Eliminating A between Eqs. (2) and (6), we get

$$\omega^2=-\ddot{x}/x, \tag{7}$$

and eliminating ω between Eqs. (2), (5), and (6), we get

$$A^2=x^2-x\dot{x}^2/\ddot{x}. \tag{8}$$

If the solution, Eq. (2), is to be subject to the condition, Eq. (4), the desired differential equation clearly is given by Eq. (3).

The solutions to Eq. (3) given by Eq. (2) and subject to Eq. (4) represent conservative motions, for writing

$$T=\tfrac{1}{2}\dot{x}^2, \tag{9}$$

$$V=-\tfrac{1}{2}x\ddot{x}=\tfrac{1}{2}x^2\left(-\frac{\ddot{x}}{x}\right)=\tfrac{1}{2}\omega^2x^2, \tag{10}$$

we have, from Eq. (3),

$$F\left(\omega^2,\frac{T+V}{\tfrac{1}{2}\omega^2}\right)=0, \tag{11}$$

and, from Eq. (4), this yields

$$T+V=\tfrac{1}{2}\omega^2A^2=\text{const.} \tag{12}$$

From Eq. (10), it is seen that the systems under consideration involve what is essentially a non-linear spring in which the effective spring constant is ω^2, and thus generally a function of amplitude.

If the implicit function $F(\omega^2,A^2)=0$ may be written explicitly as

$$\omega^2=f(A^2) \tag{13}$$

or

$$A^2=g(\omega^2), \tag{14}$$

Eq. (3) takes the form

$$\ddot{x}+xf[x^2-(x\dot{x}^2/\dot{x})]=0 \tag{15}$$

or

$$(x\dot{x}^2/x)-x^2+g(-\ddot{x}/x)=0, \tag{16}$$

as the case may be.

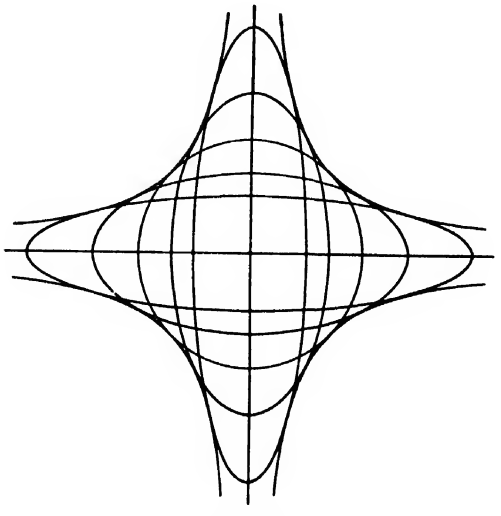


FIG. 1.

Defining the new function

$$h(z)=zg(z) \tag{17}$$

permits writing Eq. (16) in the more convenient form

$$x\ddot{x}-\dot{x}^2-h(\ddot{x}/x)=0. \tag{18}$$

From Eqs. (13) and (14), it is observed that f and g are non-negative functions which need be defined only for positive values of the argument. Thus, these remarks are also true of the function h.

Equation (3) as written may be of degree greater than one, and this leads to ambiguities in interpreting the physical motion. It is, in effect, equivalent to the product of one or more equations of the type

$$\dot{x}+x\psi_i(x,\dot{x})=0 \tag{19}$$

each of first degree and capable of physical interpretation.

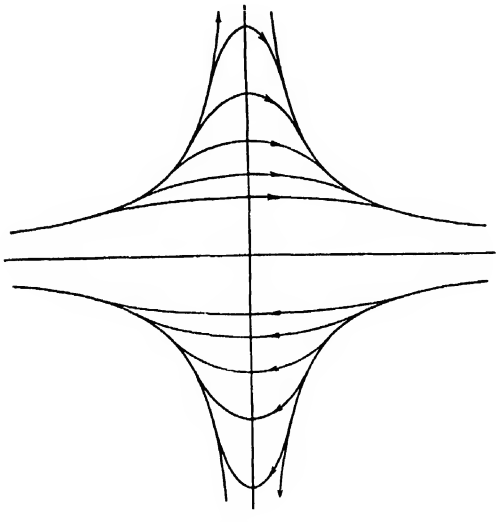


FIG. 2.

From Eq. (19), it is seen that

$$\psi=-\dot{x}/x \tag{20}$$

and

$$x^2+(\dot{x}^2/\psi)=x^2-(xx^2/x). \tag{21}$$

Thus,

$$F[\psi,x^2+(\dot{x}^2/\psi)]=0 \tag{22}$$

can be solved for

$$\psi=\psi_i(x,\dot{x}), \tag{23}$$

where the subscript i indicates that there may be several solutions.

Further points of interest will be forthcoming from examination of two simple examples. For convenience, the possible physical motions are discussed by reference to their trajectories in the displacement-velocity

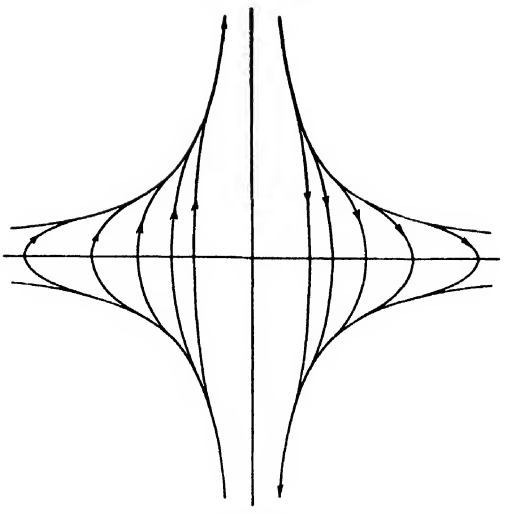


FIG. 3.

"phase" plane.[1] In these discussions, the notation

$$y=\dot{x} \tag{24}$$

will be used.

EXAMPLE 1

Let the relation between frequency and amplitude be given by

$$\omega^2=M^2/A^4, \quad (M>0). \tag{25}$$

Equation (3) becomes

$$\frac{\dot{x}}{x}+\frac{M^2}{(x^2-x\dot{x}^2/x)^2}=0, \tag{26}$$

with a solution

$$x=A\sin[(Mt/A^2)+\phi] \tag{27}$$

for which the phase trajectories are given by

$$(x^2/A^2)+(A^2y^2/M^2)=1. \tag{28}$$

The solution represents simple harmonic motion and the trajectories are concentric ellipses each of area πM. The trajectories are shown in Fig. 1. These ellipses are tangent to the equilateral hyperbolas

$$4x^2y^2=M^2 \tag{29}$$

at the points

$$\left.\begin{matrix}x=\pm(A\sqrt{2})/2\\ y=\pm(M\sqrt{2})/2A\end{matrix}\right\}. \tag{30}$$

Through every point for which $4x^2y^2<M^2$ there pass two ellipses and there are obvious difficulties in attempting to interpret the physical motion.

[1] For a discussion of and examples illustrating this type of representation, cf., N. Minorski, *Introduction to Non-Linear Mechanics* (J. W. Edwards, Ann Arbor, Michigan, 1947), especially Chapters 1–3.

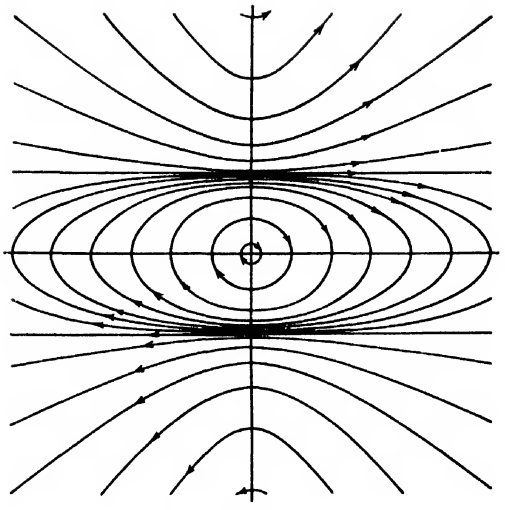


FIG. 4.

In this case, Eq. (22) becomes

$$\psi=\frac{M^2}{(x^2+y^2/\psi)^2} \tag{31}$$

from which we obtain the following two equations in the form of Eq. (19):

$$\ddot{x}+\frac{1}{x^3}\left\{\frac{M^2}{2}-x^2\dot{x}^2-M\left(\frac{M^2}{4}-x^2\dot{x}^2\right)^{\frac{1}{2}}\right\}=0 \tag{32}$$

$$\ddot{x}+\frac{1}{x^3}\left\{\frac{M^2}{2}-x^2\dot{x}^2+M\left(\frac{M^2}{4}-x^2\dot{x}^2\right)^{\frac{1}{2}}\right\}=0. \tag{33}$$

These two equations are, of course, also obtained by solving Eq. (26) for $\ddot{x}$. In each case the accelerations are real only if

$$4x^2y^2\leqq M^2, \tag{34}$$

which explains why no trajectories are found outside the region bounded by the hyperbolas of Eq. (29).

In the case of Eq. (32), a solution is given by Eq. (27), and this solution is valid only for

$$x^2\leqq A^2/2, \tag{35}$$

in other words, only for those values of x between the points where the elliptical trajectory (given by Eq. (28)) becomes tangent to the hyperbolas of Eq. (29), (cf. Eq. (30)).

In the case of Eq. (33), the solution afforded by Eq. (27) is valid only for

$$x^2\geqq A^2/2. \tag{36}$$

In both cases, the hyperbolic envelope also represents a solution valid for all x, the solution being

$$x^2=C\pm Mt, \tag{37}$$

the corresponding trajectories being the hyperbolas themselves (Eq. (29)).

Figures 2 and 3 show the trajectories corresponding to Eq. (32) and Eq. (33), respectively. It is seen that Fig. 1 represents a superposition of these.

In the cases of Eq. (32), by writing

$$\begin{cases} y=\dot{x}=F_1(t;x,y) \\ \dot{y}=-(1/x^3)\{(M^2/2)-x^2y^2 \\ \qquad -M[(M^2/4)-x^2y^2]^{\frac{1}{2}}\}=F_2(t;x,y) \end{cases} \tag{38}$$

we have a system of two first-order equations to which the ordinary test for uniqueness applies.[2] It is clear that both functions F_1 and F_2 satisfy Lipschitz conditions

$$|F_i(t;x_1,y_1)-F_i(t;x_2,y_2)|<k(|x_1-x_2|+|y_1-y_2|) \tag{39}$$

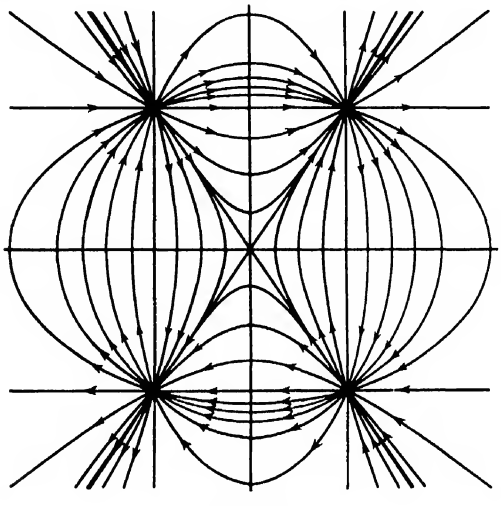


FIG. 5.

in any domain D which does not include any part of the y axis or any part of the hyperbolas of Eq. (29). Since (real) solutions are defined only for $4x^2y^2\leqq M^2$, we can assert their uniqueness except possibly for $x=0$, and upon the hyperbolas.

The trajectories of Fig. 2 show that indeed the solution is not unique on the hyperbolas. However, a unique solution does exist for any set of initial conditions corresponding to a point on the interior of the region bounded by the hyperbolas. For representative points on the y-axis, the situation must be investigated by the following limiting process:

$$\begin{aligned} \lim_{x\to 0}\dot{x}&=\lim_{x\to 0}\left[-\frac{1}{x^3}\left\{(M^2/2)-x^2y^2 \right.\right. \\ &\qquad \left.\left. -M[(M^2/4)-x^2y^2]^{\frac{1}{2}}\right\}\right] \\ &=\lim_{x\to 0}\left[-(xy^4/M^2)+0(x^3y^6/M^4)\right]=0. \end{aligned} \tag{40}$$

(This result holds regardless of the value of y.) Thus, a particle starting with any velocity and zero displace-

[2] Cf. E. L. Ince, *Ordinary Differential Equations* (Dover Publications, New York, 1944), p. 71.

ment continues with the same velocity until the representative point moves off the y-axis into the region to which the preceding uniqueness test applies. The origin provides an exception to this statement, but it is clear that in this case the particle will remain with zero velocity and zero displacement.

The trajectories of Fig. 2 do not clearly show the behavior if the initial conditions involve $y_0=0$. In such a case, the acceleration vanishes identically and the solutions is $x\equiv x_0$; such a state of equilibrium is evidently unstable since if any positive (negative) velocity is given to the particle, its displacement will tend to positive (negative) infinity.

A somewhat similar analysis may be applied to Eq. (33). This will not be given here in full detail, but it may be remarked that the motion is not defined if the initial conditions involve zero displacement, i.e., the representative point on the y-axis.

Subject to the exceptions noted in the two preceding paragraphs, it is seen that, if the initial conditions are such that

$$4x_0^2\dot{x}_0^2<M^2, \tag{41}$$

the motion will be simple harmonic until the repre-

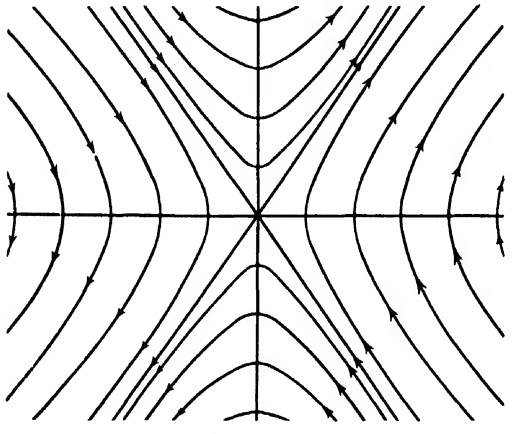


Fig. 6.

sentative point in the phase plane, moving along an elliptical path, touches one of the bounding hyperbolas. From this time on, the point follows the hyperbola and the motion is described by Eq. (37). There is no ambiguity unless the representative point for the initial conditions lies upon the bounding hyperbola in either the second or fourth quadrants. In such a case there is an instability and the representative point may follow a hyperbola or break away and follow an ellipse at any time. It may be remarked that hyperbolic trajectories have some properties similar to those of "limit cycles."

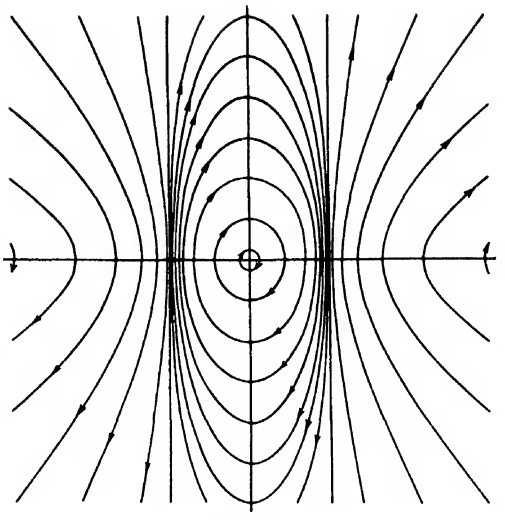


Fig. 7.

EXAMPLE 2

A fairly general case is represented by the condition

$$\omega^2=\frac{aA^2+b}{cA^2+d}, \tag{42}$$

where a, b, c, and d are constants such that $ad\neq bc$. We are led to the equations

$$\ddot{x}+x\left\{\frac{(ax^2+b-c\dot{x}^2)\pm[(ax^2+b-cx^2)^2+4a\dot{x}^2(cx^2+d)]^{\frac{1}{2}}}{2(cx^2+d)}\right\}. \tag{43}$$

The function

$$x=A\sin\left[\left(\frac{aA^2+b}{cA^2+d}\right)^{\frac{1}{2}}t+\omega\right] \tag{44}$$

is a solution to Eq. (43) with the plus sign if and only if

$$\frac{acA^4+[bc\sin^2\theta+ad(1+\cos^2\theta)]A^2+bd}{(cA^2+d)}<0, \tag{45}$$

where

$$\theta=\left(\frac{aA^2+b}{cA^2+d}\right)^{\frac{1}{2}}t+\phi. \tag{46}$$

Similarly, Eq. (44) is a solution to Eq. (43) with the minus sign if and only if the inequality opposite to that of condition (45) is satisfied.

This example is too involved to work out in full detail here, so we consider the special case where $a=0$. With this simplification we are led to:

$$[x^2+(d/c)]\ddot{x}/x=\dot{x}^2-(b/c). \tag{47}$$

In this case the investigation naturally divides itself into four cases.

Case (1). $(d/c)>0$; $(b/c)>0$

If $\dot{x}^2<(b/c)$, a solution is

$$x=A\sin\left[\left(\frac{b}{cA^2+d}\right)^{\frac{1}{2}}t+\phi\right] \tag{48}$$

with trajectories

$$x^2/A^2+[(cA^2+d)/(bA^2)]y^2=1. \tag{49}$$

If $\dot{x}^2>(b/c)$, a solution is

$$x=A\sinh\left[\left(\frac{b}{cA^2-d}\right)^{\frac{1}{2}}t+\phi\right] \tag{50}$$

with trajectories

$$[y^2(cA^2-d)/bA^2]-(x^2/A^2)=1. \tag{51}$$

For the whole phase plane the trajectories are shown in Fig. 4.

Case (2). $(d/c)<0$; $(b/c)>0$

In this case if $A^2<-(d/c)$, the frequency becomes imaginary and we cannot expect simple harmonic motion. The solution and trajectories given by Eqs. (48) and (49) hold for $-(d/c)<A^2<\infty$. For $A^2<-(d/c)$, the solutions and trajectories of Eqs. (50) and (51) hold as well as those given by

$$x=A\cosh\left[\left(\frac{-b}{cA^2+d}\right)^{\frac{1}{2}}t+\phi\right] \tag{52}$$

and

$$(x^2/A^2)+(y^2(cA^2+d)/A^2b)=1. \tag{53}$$

All the trajectories pass through the points $x=\pm(-d/c)^{\frac{1}{2}}$, $y=\pm(b/c)^{\frac{1}{2}}$, which represent instability in the sense that many motions of the representative point are possible at these points. The trajectories are shown in Fig. 5.

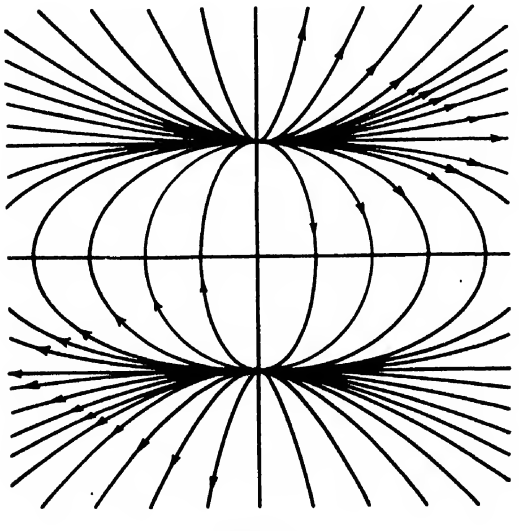


FIG. 8.

Case (3). $(d/c)>0$; $(b/c)<0$

In this case the frequency is imaginary and there can be no periodic motions as may be seen directly from Eq. (47). The singularity at $x=\dot{x}=0$ is an ordinary saddle point. The solutions and corresponding trajectories are given by Eqs. (50), (51), (52), and (53). The trajectories are shown in Fig. 6.

Case (4). $(d/c)<0$; $(b/c)<0$

If $A^2<-(d/c)$, Eqs. (48) and (49) give the solution and trajectories respectively. If $A^2>-(d/c)$, Eqs. (52) and (53) give these results. The trajectories are shown in Fig. 7.

Two interesting special cases are obtainable from this analysis. If $b=c$, and $d=0$, we have the equation

$$\dot{x}^2-x\ddot{x}=1 \tag{54}$$

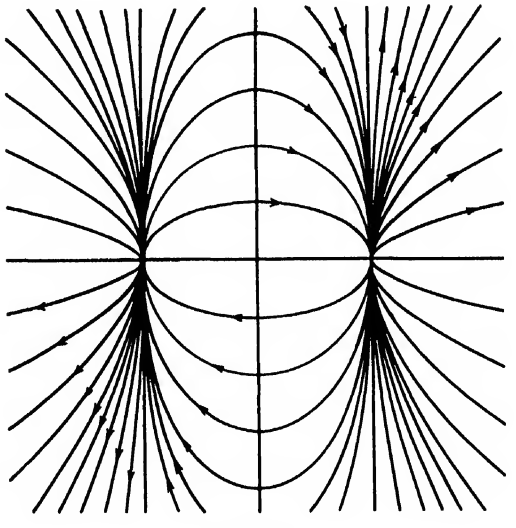


FIG. 9.

for which, if $x^2\gtrless 1$, the solution is

$$x=A\sin[(t/A)+\phi] \tag{55}$$

and the trajectories are

$$(x^2/A^2)+y^2=1. \tag{56}$$

If $\dot{x}^2\gtrless 1$, the solution is

$$x=A\sinh[(t/A)+\phi] \tag{57}$$

and the trajectories are

$$y^2-(x^2/A^2)=1. \tag{58}$$

All trajectories pass through the points $x=0$; $y=\pm 1$. (Fig. 8).

If $b=0$, and $d=-c$, we have the equation

$$\ddot{x}(1-x^2)+x\dot{x}^2=0. \tag{59}$$

If $x^2 \leqq 1$, the solution is given by

$$x = \sin[(t/A)+\phi] \tag{60}$$

and the trajectories by

$$x^2 + A^2y^2 = 1. \tag{61}$$

If $x^2 \geqq 1$, the solution is given by

$$x = \cosh[(t/A)+\phi] \tag{62}$$

and the trajectories by

$$x^2 - A^2y^2 = 1. \tag{63}$$

The trajectories are shown in Fig. 9.

The ordinary Lipschitz condition[2] for (existence and) uniqueness of the solutions suffices to demonstrate uniqueness for solutions to Eq. (47) except for $x = \pm(-d/c)^{\frac{1}{2}}$. The implications of this are apparent in Cases (2) and (4). In Cases (1) and (3) there is no exception to uniqueness.

The Analogy between Hydraulic Jumps in Liquids and Shock Waves in Gases*

F. R. Gilmore, M. S. Plesset, and H. E Crossley, Jr.
Hydrodynamics Laboratory, California Institute of Technology, Pasadena, California
(Received September 30, 1949)

The theory of the hydraulic jump is presented briefly, and the analogy between this phenomenon and the compression shock wave in gases is pointed out. The results of experimental measurements of hydraulic-jump intersections on a water table are reported. Considerable disagreement between theory and experiment is found. Other investigators have noted a disagreement between theory and experiment for compression shock intersections in gases. The discrepancy in the aerodynamic case appears unlike that found in the hydraulic case. Possible reasons for the discrepancy in the hydraulic case are discussed; some sources of error are peculiar to hydraulic jumps and do not apply to compression shocks. Such factors limit the utility of the water table as an analog device.

INTRODUCTION

IT has been known for a long time that liquid flow in open channels is of two different types: relatively smooth "streaming" flow at low speeds, and high speed "shooting" flow characterized by standing waves and frequently by sudden changes in depth known as hydraulic jumps. More recently, the increased study of compressible-flow phenomena has led investigators to note the resemblances between streaming channel flow and subsonic compressible flow, and between shooting channel flow and supersonic compressible flow. This similarity was put on a mathematical basis by Jouguet[1] and Riabouchinsky.[2] Further theoretical investigations have been carried out by von Kármán[3] and by Preiswerk[4]; experimental studies have been made by Ippen[5] and by Binnie and Hooker.[6] Later work[7] has been concerned with the application of the theory to model testing, except for a recent theoretical paper by Stoker.[8]

The analogy between liquid flow with a free surface and two-dimensional gas flow has been found useful for qualitative investigations of high speed gas flow by means of relatively simple water-table installations. However, recent advances in aerodynamics have brought many problems to the point where more accurate information is required than can be obtained from the hydraulic analogy. The present utility of the water table as an analog device is limited to special problems in aerodynamics, such as transient phenomena in high speed flow, and shock-wave intersection phenomena, particularly of the Mach type.

The discussion here is concerned with the analogy between hydraulic-jump intersections on a water table and shock-wave intersections in gases.

THEORY

Shock Waves in Compressible Gas Flow

As is well known, a compression wave of finite amplitude in a gas grows continually steeper as the wave progresses until a finite limiting steepness determined by viscous and heat-conduction effects is attained. The limiting thickness of the shock in the usual situation is negligible compared with the other dimensions in the flow field. The flow relations across a straight compression shock are determined by the conservation laws for mass, momentum, and energy without consideration of the conditions inside the shock. The results are the

* This study was supported by the ONR and the Navy Bureau of Ordnance.

[1] E. Jouguet, J. de math. pures et appliq. (Series 8) **3**, 1 (1920).
[2] D. Riabouchinsky, Comptes Rendus **195**, 998 (1932); **199**, 632 (1934); **202**, 1725 (1936).
[3] Th. von Kármán, Zeits. f. angew. Math. Mech. **18**, 49 (1938).
[4] E. Preiswerk, N.A.C.A. Tech. Memo. 934 and 935 (1940); translated from Mitteilungen Inst. f. Aerodyn., No. 7, Eidg. Tech. Hochsch., Zurich (1938).
[5] A. Ippen, Ph.D. thesis, California Institute of Technology (1936).
[6] A. M. Binnie and S. G. Hooker, Proc. Roy. Soc. **159**, 592 (1937).
[7] See, for example, Orlin, Lindner, and Bitterly, N.A.C.A. Report 875 or Tech. Note 1185 (1947).
[8] J. J. Stoker, Commun. App. Math. **1**, 1 (1948).

familiar Rankine-Hugoniot equations.[9] Figure 1 shows curves of the density ratio, temperature ratio, and the square root of the pressure ratio across a normal shock as functions of the initial Mach number for perfect gases having specific-heat ratios of $\gamma=1.4$ and $\gamma=2$.

The flow relations across a straight shock which is oblique to the flow may be obtained by superimposing on the normal velocity an additional velocity tangential to the shock which remains unchanged across the shock. For a curved shock, if the radius of curvature is much larger than the shock thickness, the flow relations may be obtained by approximating the curved shock with a number of straight segments.

Hydraulic Jumps in Liquid Flow with a Free Surface

If an elevation wave of finite amplitude is produced on the surface of a liquid, Stoker[9] has shown from the equations of motion that the wave front will grow continually steeper as it progresses, until finally the wave leans forward. At this point experiments show that the wave form deviates from the theoretical form, with the formation of a breaker or "roller" and the establishment of a turbulent wave of constant shape (except for minor fluctuations of short period). Such a steady finite wave is called a hydraulic jump. Bakhmeteff and Matzke[10] have found experimentally that

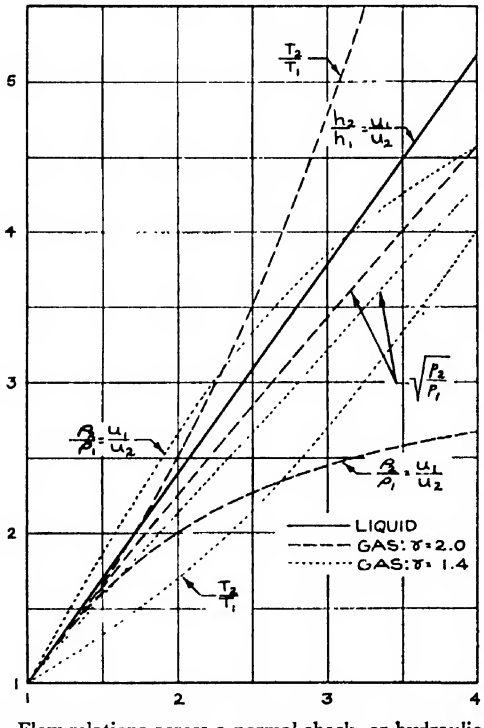


FIG. 1. Flow relations across a normal shock, or hydraulic jump. M_1 = Mach number ahead of shock.

[9] See R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves* (Interscience Publishers, Inc., New York, 1948), pp. 116–125, or similar texts on compressible flow.

[10] B. Bakhmeteff and A. Matzke, Trans. A.S.C.E. **101**, 630 (1936).

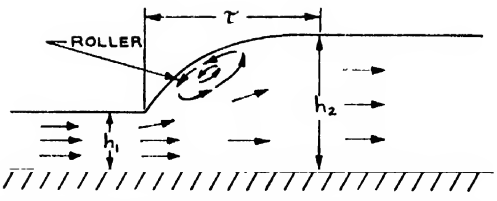


FIG. 2. Typical profile of a strong hydraulic jump ($h_2/h_1>2$).

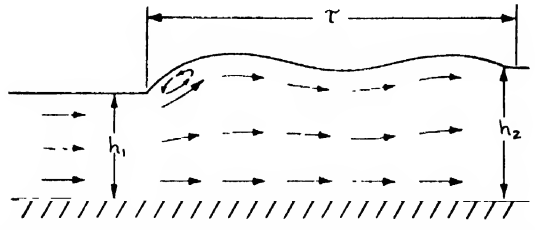


FIG. 3. Typical profile of a weak hydraulic jump ($h_2/h_1<2$).

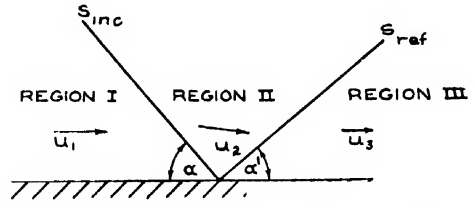


FIG. 4. Regular reflection of a shock from a rigid wall.

strong hydraulic jumps with a height ratio, h_2/h_1, greater than 2 have the simple profile shown in Fig. 2; in these strong hydraulic jumps, the thickness, τ, of the jump is four to six times the height of the jump, h_2-h_1. Weaker jumps generally have the undulatory form shown in Fig. 3. The surface first rises above the final level and then oscillates with diminishing amplitude about the final level. The thickness of the undulatory jump is not well defined, but the undulations generally die out at distances of three to ten times the initial water height, h_1, the distance being greater for the smaller jumps. A theoretical explanation of this general behavior has not as yet been given.

As in the case of compression shocks in gases, flow relations across straight hydraulic jumps are determined by the conservation laws for mass, momentum, and energy without consideration of the conditions inside the jump. The result for the height and velocity ratio of a jump normal to the flow is

$$h_2/h_1=u_1/u_2=(2u_1^2/gh_1+\tfrac{1}{4})^{\frac{1}{2}}-\tfrac{1}{2}. \qquad (1)$$

The height ratio given by Eq. (1) is plotted against initial Mach number as the solid curve in Fig. 1. "Mach number" in this case is defined as $u/(gh)^{\frac{1}{2}}$, since $(gh)^{\frac{1}{2}}$ is the velocity of infinitesimal surface waves.

Flow relations across oblique hydraulic jumps may be calculated in the same way as for oblique compression shocks. Similarly, flow relations across curved jumps can be found, provided the radius of curvature is much larger than the jump thickness. In many experi-

mental situations, however, the thickness of the jump may be of the same order of magnitude as the radius of curvature. Existing theory does not cover this case.

The Analogy between Shocks and Jumps

It may be shown[1–4] that the smooth free-surface flow of a liquid over a horizontal bed obeys equations of motion similar to those for the isentropic two-dimensional flow of a hypothetical gas with specific heat ratio $\gamma=2$. The mathematical correspondence is between the following quantities: the liquid velocity ratio and the gas velocity ratio; and the liquid depth ratio on the one hand and the gas density ratio, the gas temperature ratio, and the square root of the gas pressure ratio, on the other hand. It is to be noted that the isentropic condition makes the gas density ratio, the gas temperature ratio and the square root of the gas pressure ratio all equal when $\gamma=2$.

Flows with shocks or jumps are not isentropic and it might be expected that the quantitative correspondence breaks down. The entropy change, however, is of the third order in the shock strength so that weak shocks are very nearly isentropic; and the quantitative similarity should then hold to a good approximation. This situation is demonstrated in Fig. 1, which compares values of analogous quantities for the compression shock and the hydraulic jump. The curves for a gas with $\gamma=2$ approach the liquid curve asymptotically as the shock becomes weaker.

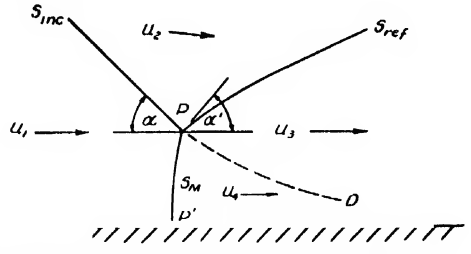


FIG. 5. Mach reflection of a shock from a rigid wall.

An analogy based on a hypothetical gas with $\gamma=2$ is somewhat unrealistic since kinetic theory shows that gases must have specific-heat ratios in the range $1<\gamma\leqslant 5/3$. Figure 1 includes curves for a gas with $\gamma=1.4$, which is the value for air. Although these curves do not have an asymptotic approach to the liquid curve as the shock strength is decreased, some of the flow parameters, e.g., u_1/u_2, have a smaller deviation over a range of strong shocks than the corresponding parameters for $\gamma=2$.

Intersections and Reflections of Hydraulic Jumps and Shock Waves

The following discussion will apply both to hydraulic jumps and compression shock waves, and the term "shock" will be used to signify either phenomenon.

Figure 4 shows the reflection of a straight shock from a plane rigid wall; it could as well refer to the intersection of two equal straight shocks where the wall becomes the plane of symmetry. The fluid in Region I, ahead of the shocks, is assumed to be either at rest or in a state of uniform motion. It is convenient to introduce a coordinate system relative to which the shocks and their intersection are stationary; in this coordinate system let u_1 be the velocity with which the fluid in Region I approaches the incident shock, S_{inc}. As the flow passes through this shock, it is slowed down and deflected to a velocity u_2 in Region II. The reflected shock, S_{ref}, must deflect the flow from the velocity u_2 to a velocity u_3 in Region III which, like u_1, must be parallel to the boundary wall or to the plane of symmetry. This condition is sufficient to determine the angle and strength of the reflected shock.[11]

When the incident shock is too strong, and when it makes too large an angle α with the wall, there is no real solution for the angle and strength of the reflected

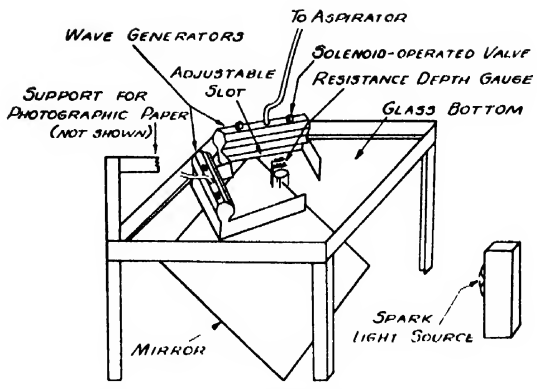


FIG. 6. A schematic representation of the experimental apparatus.

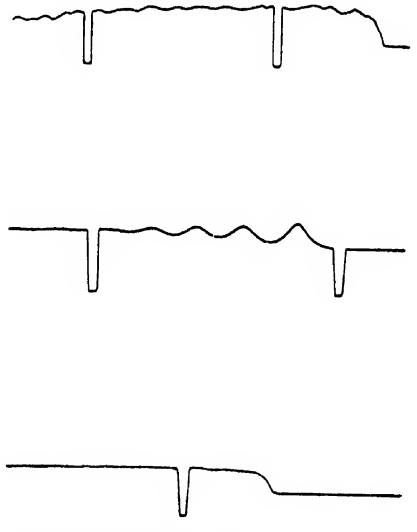


FIG. 7. Typical hydraulic-jump forms. Top curve: Strong jump with irregular roughness. Center curve: Weak jump with undulations. Bottom curve: Smooth jump of medium strength. The breaks in these records, which give the reference zero for depth, are spaced 0.6 sec. apart.

[11] See reference 9, pp. 318–350.

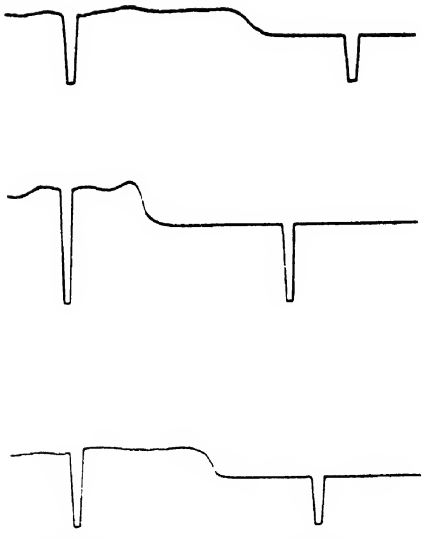


FIG. 8. Effects of surface-tension reducing agents on the jump form. Top curve: Without detergent. Center curve: Strong detergent (lauryl isoquinolinium bromide 0.1 percent). Bottom curve: Weak detergent (Kodak Photo-Flow, 0.5 percent)

shock so that the "regular" reflection scheme shown in Fig. 4 becomes impossible. In this situation the actual flow is found experimentally to follow the scheme shown in Fig. 5. A third shock, the "Mach," S_M, appears, and its length, i.e., the distance PP', increases with time at a fairly constant rate. The Mach, S_M, is frequently curved, and also the reflected shock, S_{ref}, is often curved near the triple-shock intersection.

The simplest theory of the Mach reflection assumes that all three shocks are straight.[11] Coordinates are chosen fixed with respect to the shock intersection P so that the problem is one of steady flow. In these coordinates, the wall (or plane of symmetry) is moving downward at a constant velocity equal to the component of u_1 or u_4 tangential to S_M. A line PD can be drawn separating the fluid which has passed through S_{inc} and S_{ref} from that which has passed through S_M. The gas pressure (or liquid height) must be continuous across PD since it is not a shock. However, fluid which has passed through a single large shock has lost more flow energy than that which has passed through two smaller shocks with the same total change in pressure (or height); hence the velocity and density of a gas (or the velocity of a liquid) will be less below PD than above it. Thus, PD is a slip-stream. Schlieren photographs of gas flows have shown such a density discontinuity at PD.

Quantitative relations for such three-shock intersections can be obtained by applying the oblique shock equations to each of the shocks. The results for some ranges of interest have been tabulated by Polachek and Seeger both for compression shocks[12] and for hydraulic jumps.[13]

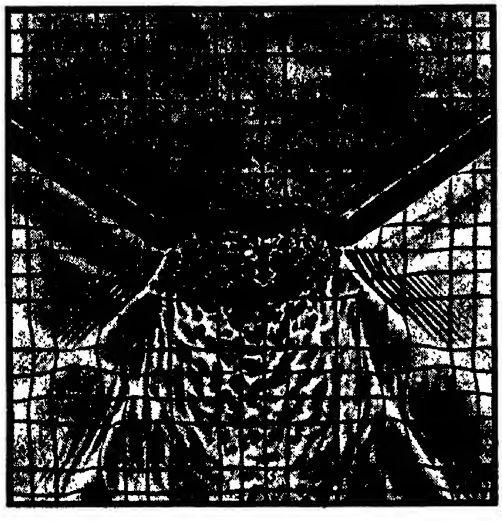


FIG. 9. Refraction photogram of medium-strength hydraulic-jump intersection ($\xi=0.45$, $\alpha=56°$).

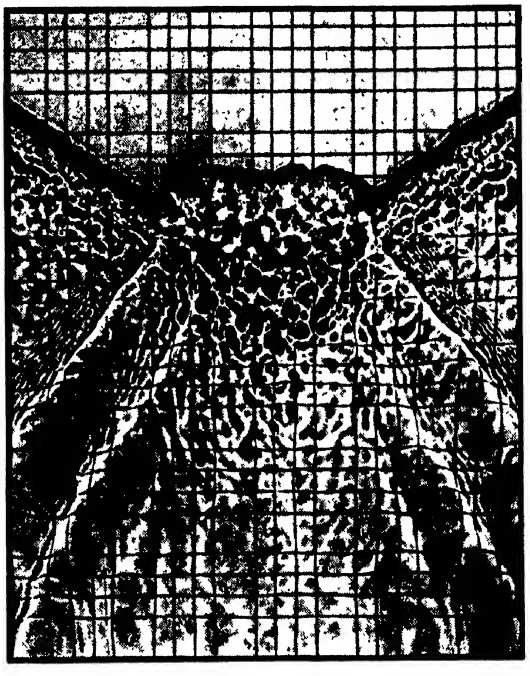


FIG. 10. Refraction photogram of strong hydraulic-jump intersection ($\xi=0.28$, $\alpha=56°$).

EXPERIMENTAL MEASUREMENTS

The purpose of the present experimental study was the investigation of regular and Mach intersections in hydraulic jumps. The apparatus used in the experiments was essentially that used by Einstein and Baird[14]

[12] H. Polachek and R. J. Seeger, Explosives Research Report No. 13, Navy Department, BuOrd, Washington, D.C. (February, 1944).

[13] H. Polachek and R. J. Seeger, Explosives Research Report No. 14, Navy Department, BuOrd, Washington, D.C. (August, 1944).

[14] H. A. Einstein and E. G. Baird, Progress Reports on the Analogy Between Surface Shock Waves on Liquids and Shocks in

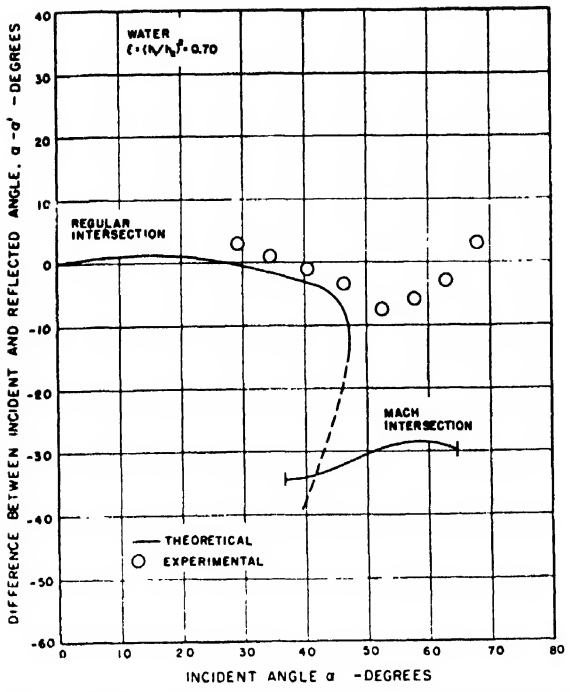


FIG. 11. Relationship between incidence and reflection angles for the intersection of weak hydraulic jumps ($\xi=0.70$).

in earlier investigations and is shown schematically in Fig. 6. The water table was approximately 5 ft. long, 4 ft. wide, and 2 in. deep. The liquid used in the tank was a 0.001 to 0.002 normal solution of manganous chloride in distilled water; the manganous chloride was added to the water so that the electrical conductivity of the solution would be more uniform. The hydraulic jumps were produced by two wave generators which were reservoirs into which a volume of water could be raised above the free surface; this volume of water could be discharged quickly through an adjustable horizontal slot 24 in. long located just below the free surface. Instantaneous water depths at any point on the water table could be determined by measuring the resistance between a pair of partially immersed platinum electrodes. A 1000-cycle alternating voltage was applied across the electrodes and the current which flowed was recorded by an oscillograph with an upper limit of about 100 cycles on its frequency response.

Typical wave depth records are shown in Fig. 7; the record is interrupted at intervals to give a zero reference line. The horizontal coordinate in these records is a time scale not a spatial scale, but a good estimate of the instantaneous wave form in space may be obtained by multiplying this time scale by the wave velocity.

In addition to a continuous record of the water depth at a point, an instantaneous refraction photogram of the entire field could be made. The bottom of the water tank was made of glass so that the light flash from a spark discharge across a $\frac{1}{8}$-in. gap could be transmitted upward through the water and the resulting refraction pattern could be recorded on photographic paper placed above the water surface.

RESULTS AND DISCUSSION

Measurements were first made to determine the shape and velocity of single hydraulic jumps produced with one wave generator alone; depths were determined with three pairs of electrodes. The initial water depth, h_1, ranged from about 0.5 cm for strong jumps to about 2 cm for weak jumps. Weak jumps were found to have an undulatory profile, stronger jumps had a smoother profile, and very strong jumps showed appreciable irregular roughness (cf. Fig. 7). These observations agree qualitatively with those of Bahkmeteff and Matzke[10] on hydraulic jumps larger by an order of magnitude than those reported here. The velocities of the hydraulic jumps were also measured and compared with the usual formula

$$u_1=\{g(h_1+h_2)h_2/(2h_1)\}^{\frac{1}{2}}, \qquad (2)$$

where h_1 is the water depth ahead of the jump and h_2 is the water depth behind the jump. The measured velocities of weak jumps agreed with the theoretical values within the experimental error of approximately two percent, but the measured velocities of the stronger jumps were up to five percent larger than the theoretical values. A possible explanation of this discrepancy is that the water released from the wave generator may tend to override the surface of the still water. The water

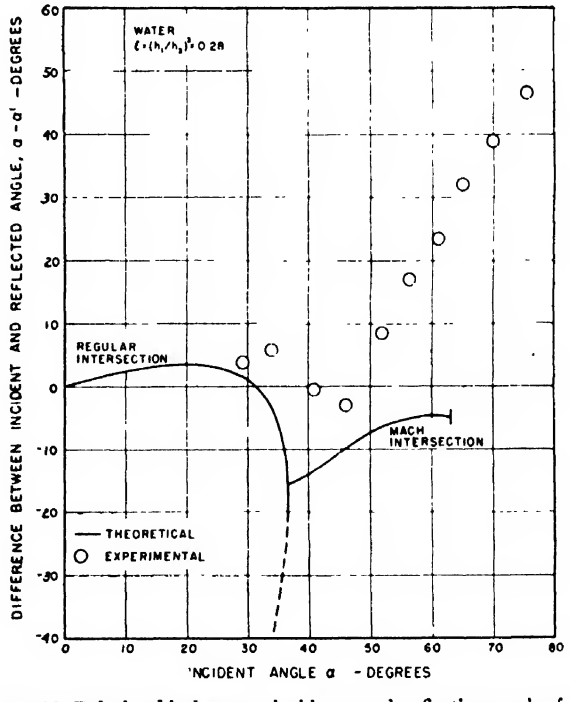


FIG. 12. Relationship between incidence and reflection angles for the intersection of strong hydraulic jumps ($\xi=0.28$).

Compressible Gases, Hydrodynamic Laboratories, California Institute of Technology (September 15, 1946 and July 30, 1947).

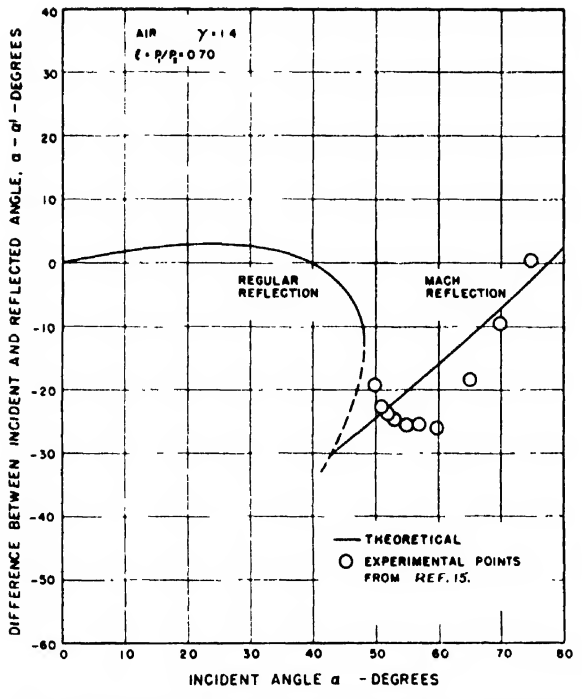


FIG. 13. Relationship between angles of incidence and reflection for the reflection of weak compression shocks ($\xi=0.70$) in air. The experimental points are from the data of Harrison and Bleakney (see reference 15).

in the jump would then have a higher velocity at the free surface than at the bed so that the simple theory would not apply.

A study was also made of the effect of the addition of surface-tension reducing agents to the water. When 0.1 percent by volume of lauryl isoquinolinium bromide was added to the water, the wave fronts formed became steeper, and the strong jumps became smoother. However, an appreciable amount of foam was produced which interfered with the photographic procedure. It was noted, moreover, that the amplitude of the undulations in the weaker jumps was increased. A weaker detergent solution, which consisted of 0.5 percent by volume of Kodak Photo-Flo (an aerosol solution), was found to give an intermediate effect: the undulations were enhanced only in the weakest waves, while the wave form of the strong waves was smoothed and steepened. The effect of these detergent solutions is shown in Fig. 8.

Measurements were made on the hydraulic jump intersections produced by two wave generators set at various angles. The jump strengths were determined by the resistance depth gauges while the general flow pattern was recorded on the refraction photograms (cf. Figs. 9 and 10).

Most of the measurements were made with jump strengths, $\xi\equiv(h_1/h_2)^2$, of 0.28, 0.45, and 0.70. For the medium and strong jumps ($\xi=0.45$ and 0.28), the aerosol solution was used, while for the weak jumps ($\xi=0.70$) detergent-free water was used, since this gave the best wave form. Figures 11 and 12 show the experimentally measured differences between the angle of incidence, α, and the angle of reflection, α', plotted against α for weak and strong jumps, respectively. Theoretical curves for the regular and Mach intersections are also shown. The agreement between theory and experiment is fair for the limited number of regular intersections investigated, but is definitely poor for the Mach intersections. There is little evidence of the sudden change in reflection angle which is predicted by the theory in passing from regular to Mach intersections.

These figures may be compared with Figs. 13 and 14, which give the corresponding data for intersections of compression shocks in air. The experimental points are taken from a report by Harrison and Bleakney.[15] The agreement between theory and experiment is seen to be better here, but an appreciable discrepancy remains, especially for the weaker shocks. There appears to be little resemblance between the trend of the disagreement in the liquid and gas cases.

One possible error in the use of the simple theory in the above comparisons is the neglect of the curvature of the Mach jump. The geometric effect of this curvature can be eliminated by considering only a small region about the triple jump intersection. Such a presentation of the data is shown in Figs. 15 and 16 where ω and ω' are the local angles which the incident and

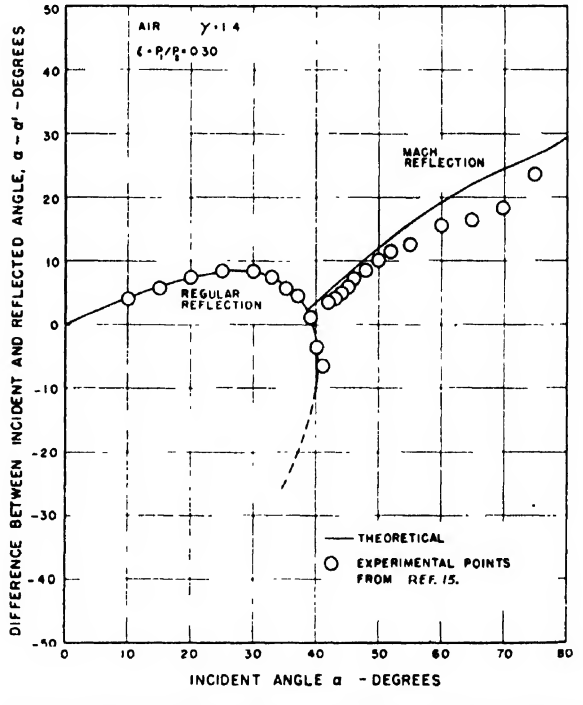


FIG. 14. Relationship between angles of incidence and reflection for the reflection of strong compression shocks ($\xi=0.30$) in air. The experimental points are from the data of Harrison and Bleakney (see reference 15).

[15] F. B. Harrison and W. Bleakney, "Remeasurement of reflection angles in regular and Mach reflection of shock waves," Physics Department, Princeton University (March, 1947).

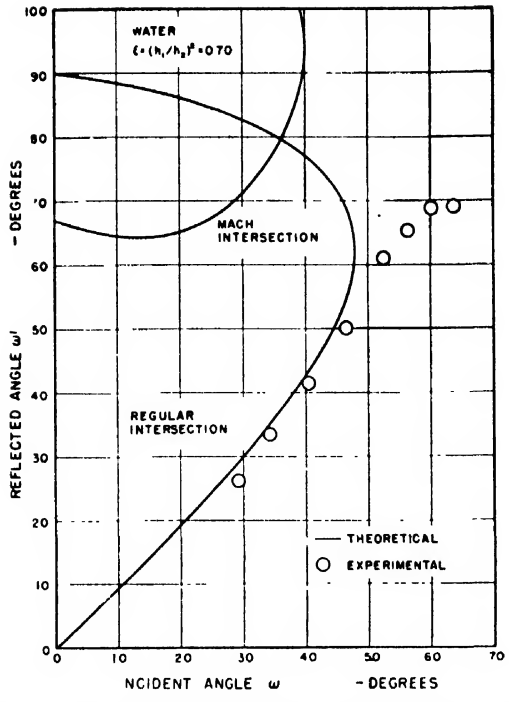


FIG. 15. Relationship between incidence and reflection angles, measured with respect to the direction of motion of the triple intersection, for the intersection of weak hydraulic jumps ($\xi=0.70$).

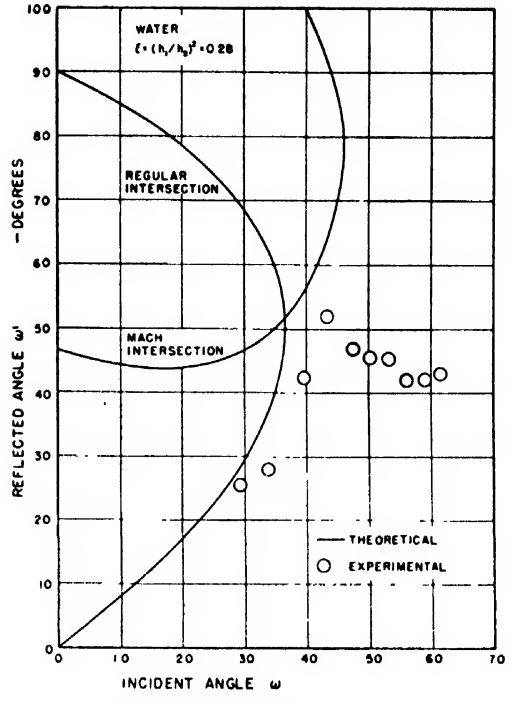


FIG. 16. Relationship between incidence and reflection angles, measured with respect to the direction of motion of the triple intersection, for the intersection of strong hydraulic jumps ($\xi=0.28$).

reflected jumps make with the direction of motion of the triple point. The weak-jump data shown in Fig. 15 show a disagreement with theory similar to that found by Harrison and Bleakney for air, although the discrepancy is somewhat greater for the hydraulic case. The strong-jump data shown in Fig. 16 show similar disagreement with theory, whereas Harrison and Bleakney found very little discrepancy in air.

Examination of the photograms indicates that the thickness of the Mach jump is not negligible compared with the radius of curvature of the Mach. The present theory does not take account of this possibility so that an error of unknown magnitude has been introduced. This difficulty does not arise in the aerodynamic case. The experimental observotions indicate an additional source of error in the simple theory: the liquid depth behind the reflected jump and the Mach jump is not uniform; in fact, there sometimes appears to be an additional hydraulic jump approximately parallel to, but behind, the reflected jump. Corresponding sources of error may exist in the aerodynamic case although direct observations on this possibility are not available.

One may conclude from the experimental observations discussed here that the deviations from the simple theory of hydraulic-jump intersections limit the use and validity of the hydraulic analog as a means for studying compression-shock intersections in gases.

An Optical Method for Measuring the Stress in Glass Bulbs

W. T. READ
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received October 6, 1949)

In the sealing and processing of electron tubes residual stresses are introduced into the glass parts. This paper formulates a practical optical method for locating and measuring the maximum tension in the walls of glass bulbs.

The basic principles of photoelasticity in three dimensions are applied to the case of a cylindrical shell. Theoretical results are verified experimentally for a tube in which the stress distribution is axially symmetrical.

Experimental results on tubes in which the stress distribution was unsymmetrical lead to the conclusion that the maximum tensile stress is proportional to the maximum retardation, the factor of proportionality being simply related to the form of the fringe pattern.

A quick and simple technique is described for making the necessary optical measurements.

The method presented herein has been used and found to be practical for routine laboratory testing of tubes in quantity.

I. INTRODUCTION

IN the sealing and processing of electron tubes, residual stresses are introduced in the bulb walls due to non-uniform heating and mismatch of thermal expansion coefficients. Failures resulting from these stresses constitute a serious problem especially in glass walled tubes. A quick and reasonably accurate method of determining the maximum stress in the bulb would permit the detection and rejection of overstressed tubes at the factory and aid in the improvement of sealing techniques.

The usual procedure is to examine the tube in a polariscope and judge the stress magnitude by observation of relative phase retardation. Because of the three-dimensional state of stress in the tube wall the quantitative conversion of the observed retardation to stress magnitude involves considerations beyond those commonly employed. It was the purpose of this investigation to discover the relation between the observed retardation and the maximum stress in the cylindrical portion of the bulb and to develop a simple laboratory procedure for making the necessary measurements.

We begin with a brief review of the stress optical relations in the general three-dimensional case.

Upon entering a doubly refracting material such as stressed glass, the incident light is resolved into two waves, plane polarized at right angles to one another and traveling with different velocities. In a material as weakly birefringent as glass the two waves may be assumed to travel in the same direction and to have the same wave normal. In Fig. 1 let the wave normal be parallel to the y axis so that the wave front is parallel to the xz plane, f being the direction of polarization of the faster moving wave and s, of the slower moving wave. We designate φ as the (variable) angle between f and the x axis. The angle φ and the relative velocities of the two waves are determined by the directions and magnitudes of the secondary principal stresses in the plane of the wave front. The axes of secondary principal stress are the normals to the two perpendicular planes through the y axis on which the shearing stress vanishes, and the normal stress is stationary. The secondary principal stresses, p and q, are respectively the maximum and minimum values of the normal stress in the xz plane.

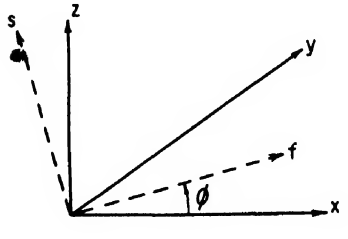


FIG. 1. Directions of polarization in plane of wave front.

We shall use the following two stress-optical laws which apply to weakly birefringent materials such as glass: (1) at each point in the stressed material the polarizing axes f and s in the plane of the wave front are along the axes of secondary principal stress, and (2) the difference between the reciprocals of the velocities of the two waves is proportional to the difference between the secondary principal stresses, i.e., to $p-q$. For most materials the direction of the maximum normal stress p is along the f axis.

The angle φ (Fig. 1) is, therefore, the angle between the direction of maximum normal stress and the x axis. Letting σ_x and σ_z be the normal components of stress in the x and z directions and τ_{xz} be the shearing component we have the relations[1]

$$\tan 2\varphi = (2\tau_{xz})/(\sigma_x - \sigma_z) \tag{1}$$

$$p - q = [(\sigma_x - \sigma_z)^2 + 4\tau_{xz}^2]^{\frac{1}{2}}. \tag{2}$$

If the light is monochromatic the vibration at any point in the glass is defined by the difference in phase ψ between the components along the f and s axes and by the ratio of amplitudes, $\tan\omega$. The two functions of y, ω and ψ, are given in terms of the functions φ and

[1] R. D. Mindlin, J. App. Phys. **10**, No. 4, 229 (April, 1939).

$p-q$ by the following two simultaneous first-order differential equations with variable coefficients:[2]

$$(d\omega/dy)+(d\varphi/dy)\cos\psi=0$$

$$\frac{\tan 2\omega}{2}\left[\frac{d\psi}{dy}+\frac{2\pi C}{\lambda}(p-q)\right]-\frac{d\varphi}{dy}\sin\psi=0, \qquad (3)$$

where λ is the wave-length of the incident light and C is the relative stress optical coefficient.

These equations were derived by F. Neumann from simple kinematical considerations involving approximations which are valid for photoelastic studies.[3] In the standard two-dimensional case the stress distribution is nearly constant through the thickness of the model so that rotation of the polarizing axes along the path of the ray may be neglected and Neumann's equations may be easily integrated to give $\omega=$ const and the standard two dimensional formula

$$\psi/2\pi=C/\lambda(p-q)d,$$

where d is the thickness of the model. In the general three-dimensional case there is rotation of the secondary principal axes and the exact solution is impractical. However, in some cases the equations can be considerably simplified subject to certain limitations on the quantities involved. Denoting by $R/2$ the ratio of rate of rotation of polarizing axes to rate of relative phase retardation referred to stationary axes and substituting into Neumann's first equation gives

$$d\psi/dy=-(2\pi C)/\lambda(p-q)[1+R\cot 2\omega\sin\psi]. \qquad (4)$$

If the rate of rotation of the secondary principal axes is small compared to the rate of relative phase retardation and if the amplitudes of the two components are equal on entry and remain approximately equal (so that ω remains close to $\pi/4$ radians) then $R\cot 2\omega$ is negligibly small compared to unity and we may use the simple formula

$$d\psi/dy=-(2\pi C)/\lambda(p-q). \qquad (5)$$

However here $p-q$ is in general an unknown function of y so that we shall leave the equation in differential form. Thus subject to the conditions on R and ω the rate of relative phase retardation depends only on the secondary principal stress difference, and not on the rotation of the polarizing axes. The exact form of the equations and the effect of large $R\cot 2\omega$ are discussed for simple states of stress in Appendix I.

II. PHOTOELASTIC THEORY OF SHELLS

In this section we shall regard the bulb as a cylindrical shell and use the concepts and approximations of shell theory. Figure 2 shows a cross section of the

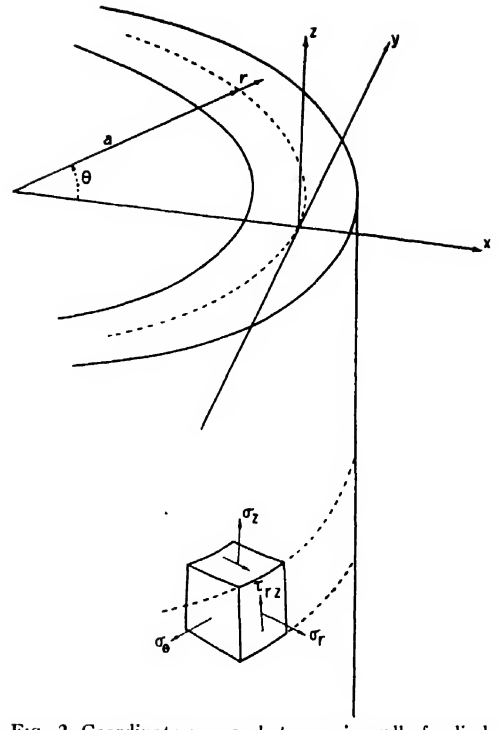


FIG. 2. Coordinate axes and stresses in wall of cylinder. Light rays are parallel to y axis.

cylinder perpendicular to the axis and a typical element of volume cut out of the tube wall. The tube is in a field of parallel rays and is surrounded by a liquid medium having the same index of refraction as the glass of the tube wall so that the light rays are not refracted on entry and exit. We designate a as the radius of the middle surface of the shell and h as the half-thickness. As before we take the y axis parallel to the light rays so that the xz plane is parallel to the wave front. We shall also use a cylindrical coordinate system constructed in the usual way with respect to the cylinder, except that the radial coordinate is measured from the middle surface of the shell.

The lower part of Fig. 3 shows the stress components in cylindrical coordinates. In shell theory the following assumptions are made regarding the stresses: the radial stress, σ_r, is identically zero, the circumferential stress, σ_θ, is constant, the axial stress, σ_z, is a linear function of r which vanishes on the middle surface, and the shearing stress τ_{rz} is a parabolic function vanishing on the inner and outer surfaces as required by the boundary conditions. The relations between the stresses σ_θ and τ_{rz} and the stresses σ_x and τ_{xz} in the plane of the wave front are

$$\tau_{xz}=\tau_{rz}\cos\theta$$
$$\sigma_x=\sigma_\theta\sin^2\theta.$$

Substituting into the formula for φ (Eq. (1))

$$\tan 2\varphi=\frac{2\tau_{rz}\cos\theta}{\sigma_\theta\sin^2\theta-\sigma_z}. \qquad (6)$$

[2] E. G. Coker and L. N. G. Filon, *Photo-Elasticity* (Cambridge University Press, London, 1931), p. 256.

[3] For the relation between Neumann's equations and Maxwell's electromagnetic equations see R. D. Mindlin and L. E. Goodman, "The optical equations of three-dimensional photoelasticity," J. App. Phys. 20, No. 1, 89–94 (January, 1949).

Since τ_{rz} must be zero on the inner and outer surfaces the angle φ between the polarizing axis of the faster moving wave and the x axis will be zero on exit and entry. Thus at the beginning and end of the path of a ray through the wall of a tube in an axially symmetrical state of stress the two polarizing axes are respectively along and perpendicular to the axis of the cylinder.

Inside the tube wall τ_{rz} will in general be non-vanishing and there will be some rotation of the secondary principal axes along the path of the ray. However, in the photoelastic study of stresses in electron tube bulbs we are interested only in the stresses at the point where failure is most likely to occur. Hence we wish to investigate the stress distribution where σ_z is a maximum and in particular to measure the maximum. Since glass is weak in tension, breakage will result from either excessive positive axial or circumferential stress. In the tubes considered here the axial stress was the critical

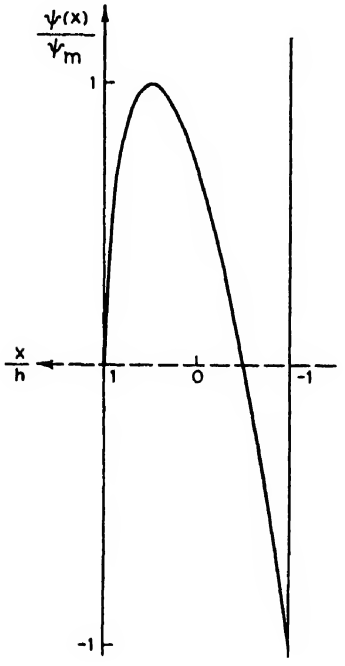


FIG. 3. Theoretical fringe pattern in bulb wall for symmetrical stress. Fringe displacement is proportional to phase difference.

one, as indicated by breakage results and by independent measurements of circumferential stress. The method of measuring σ_θ is described in Appendix II.

The axial stress is given by

$$\sigma_z=\sigma(r/h),$$

where σ is a function of z and is the maximum of σ_z for a given cross section. The absolute maximum of σ_z occurs where σ has its maximum value, and therefore where $d\sigma/dz=0$. Using this condition and the equation of equilibrium in the axial direction,[4]

$$\frac{\partial\sigma_z}{\partial z}+\frac{1}{r+a}\frac{\partial}{\partial r}(r+a)\tau_{rz}=0 \qquad (7)$$

[4] S. Timoshenko, *Theory of Elasticity* (McGraw-Hill Book Company, Inc., New York), p. 309. (Timoshenko's r is our $r+a$.)

we have $(r+a)\tau_{rz}=$constant. However, from the boundary conditions, τ_{rz} is zero at the inner and outer surfaces, so that $\tau_{rz}=0$ throughout the thickness of the wall, and from (6) $\varphi=0$ or $\pi/2$ over the cross section. We have now established the following useful result: in a cylindrical shell there is no rotation of the polarizing axes along the path of a ray lying in the cross section where the axial stress reaches its maximum value. Therefore Eq. (5) applies exactly and ω is constant and equal to the angle between the x axis and the plane of polarization of the incident light.

To discover the meaning of the observed retardation in terms of stress we use (5). In the cross section of maximum axial stress $\tau_{rz}=0$ so that (2) simplifies to

$$p-q=\sigma_x-\sigma_z=\sigma_\theta\sin^2\theta-\sigma_z.$$

Letting σ_m be the maximum value of σ_z we have

$$d\psi/dy=2\pi C/\lambda[(\sigma_m/h)r-\sigma_\theta\sin^2\theta]. \qquad (8)$$

To integrate this equation we express r and θ in terms of the variable y and the parameter x which is constant for a given ray and is equal to the minimum distance between the ray and the middle surface. When a is large compared to r and x we have the approximations:

$$y^2/2a=r-x \quad \sin\theta=y/a.$$

As the ray passes through the tube wall y varies between $-[2a(h-x)]^{\frac{1}{2}}$ and $+[2a(h-x)]^{\frac{1}{2}}$. Since the incident light is monochromatic and plane polarized there is a single λ and the phase difference on entry is zero. The phase difference on exit $\psi(x)$ is a function of the parameter x given by the integral of Eq. (8)

$$\frac{\psi(x)}{2\pi}=\frac{CL\sigma_m}{\sqrt{2}\lambda}\left[\frac{x}{h}+\frac{1}{3}\left(1-2\frac{\sigma_\theta}{\sigma_m}\frac{h}{a}\right)\left(1-\frac{x}{h}\right)\right][1-(x/h)]^{\frac{1}{2}},$$

where $L=2(4ah)^{\frac{1}{2}}=2[(\text{diameter})\times(\text{wall thickness})]^{\frac{1}{2}}$ is the length of the path of a ray tangent to the inner surface of the bulb.

In all of the tubes observed σ_θ was of the same order of magnitude as or less than σ_m, and in electron tubes $h\ll a$ so that we can neglect $2(\sigma_\theta/\sigma_m)(h/a)$ in comparison with unity. Then

$$\frac{\psi(x)}{2\pi}=\frac{CL\sigma_m}{3\sqrt{2}\lambda}\left(1+2\frac{x}{h}\right)[1-(x/h)]^{\frac{1}{2}}. \qquad (9)$$

We now have another useful result concerning cylindrical shells; namely, that in the cross section where the axial stress is a maximum the relative phase retardation depends on only the axial stress provided that the maximum axial stress is of the same order of magnitude as or greater than the circumferential stress. If σ_θ is much greater than σ_z the technique described in Appendix II should be used to measure σ_θ and the formula for σ_m connected accordingly.

Figure 3 is a plot of $[\psi(x)]/\psi_m$ as given by Eq. (9), where ψ_m is the maximum of $\psi(x)$.

Since we have assumed an axially symmetrical distribution of stress the first experiments to check the theory were made on a type of tube in which the fringe pattern remained constant as the tube was rotated about its axis. The phase difference in the cross section of maximum axial stress was measured using polarized light and a quartz wedge. Figure 4 (center fringe) shows the fringe pattern and tube wall focused in the same plane. The retardation caused by passage through the tube wall is proportional to the displacement of the fringe in the vertical direction, the distance between fringes corresponding to a retardation of 2π radians or one wave-length. Since the stress does not vary appreciably in the vertical direction in a distance equal to the maximum fringe displacement the fringe pattern itself is approximately a plot of the measured $\psi(x)$ *vs.* x.

Comparison of Figs. 3 and 4 shows that the theoretical curve of Fig. 3 derived from shell theory agrees very well with the pattern actually observed.

The formula for $\psi(x)$ Eq. (9) may be used to relate the measured maximum phase difference ψ_m to the maximum axial stress σ_m. Letting $\sigma_0=\lambda/CL$ be the stress which would produce a relative retardation of one wave-length, that is a phase difference of 2π radians, in the distance, L, traveled by the innermost ray of the tube wall, and letting $R_m=\psi_m/2\pi$ be the relative retardation in wave-lengths we have

$$\sigma_m=3\sigma_0 R_m. \tag{10}$$

If λ is expressed in angström units, L in millimeters and C in 10^{-2} brewsters than σ_0 and σ_m will be in kilograms per square millimeter.

III. ROTATION OF POLARIZING AXES — UNSYMMETRICAL STRESSES

We shall now consider in more detail the formation of the fringe pattern in Fig. 4. After emergence from the tube the originally plane polarized light is elliptically polarized. Since the f and s axes on exit are parallel to the z and x axes which in these experiments are vertical and horizontal respectively, the quartz wedge is oriented with its axes also vertical and horizontal. The relative retardation produced by the wedge varies linearly in the vertical direction so that for any value of x there is always a point on the wedge where the relative retardation caused by the tube $\psi(x)$ is equal and opposite to that caused by the wedge. At this point the light reaching the analyzer will be plane polarized at an angle $\omega(x)$ to the x axis. If the analyzer is set at an angle of $\omega(x)+(\pi/2)$ radians the light will be extinguished and the point will appear as a dark spot in the observed fringe pattern, the displacement of the point being proportional to $\psi(x)$. However this condition cannot be simultaneously satisfied for all x unless $\omega(x)$ is constant. For a variable $\omega(x)$ we would observe not a continuous fringe in the tube wall but an isolated dark spot which would move across the thickness of the tube wall as the analyzer was rotated.

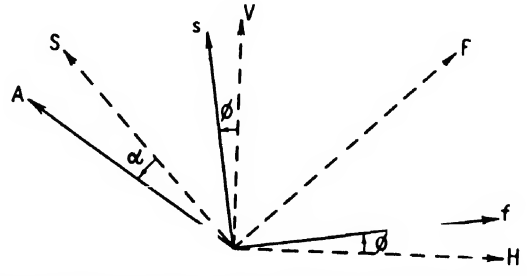


FIG. 5. Optical axes of specimen, quarter wave plate and analyzer.

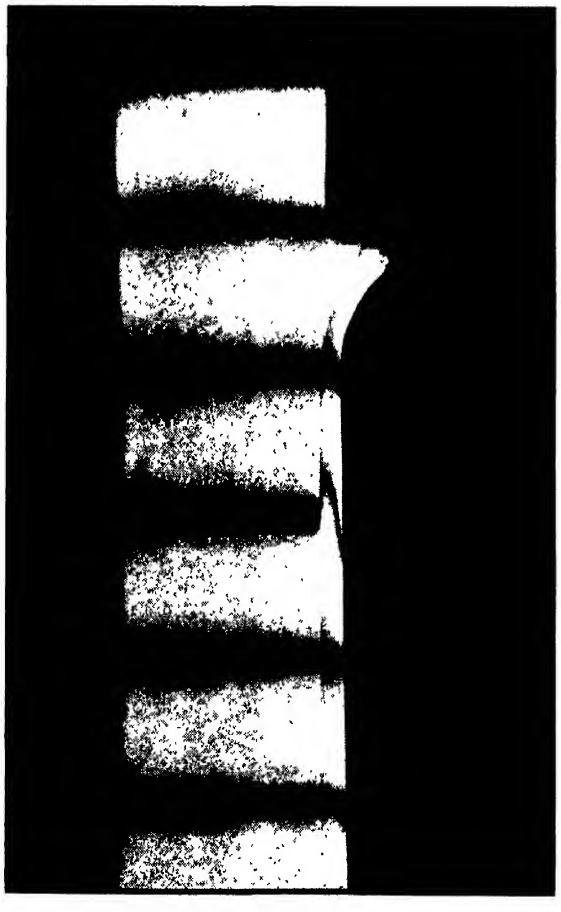


FIG. 4. Actual fringe pattern in bulb wall for symmetrical stress.

The theory of cylindrical shells predicts that where the axial stress is a maximum $\omega(x)$ will be constant and there we would expect to observe a continuous fringe. Actually, however, continuous fringes could be observed at every cross section indicating that throughout the bulb the deviation of ω from the initial value of $\pi/4$ radians is small enough to remain within the band largely extinguished by the analyzer. Rotating the analyzer obscured rather than improved the fringe pattern as a whole. Also the intensity at any point on the fringe was increased by rotating the analyzer by more than $\pm 5°$, indicating that the variation in ω is

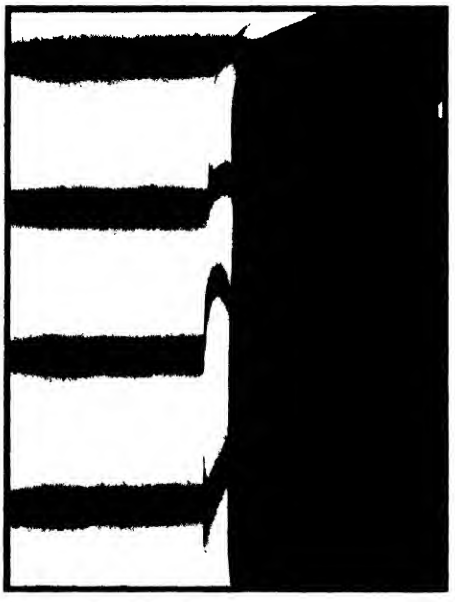


FIG. 6. Fringe pattern in wall of unsymmetrically stressed bulb. Maximum fringe displacement indicates maximum stress.

within these limits. From Eq. (3) the small variation in ω implies that the rotation of the polarizing axes is small which in turn implies that the shearing stress τ_{rz} is small compared with the axial stress σ_z. Thus the results obtained for the cross section of maximum axial stress apply approximately to every cross section and the phase difference at every level may be regarded as a function of the axial stress.

Where the stress distribution is not axially symmetrical the problem is considerably more complicated and we have the stress components $\tau_{r\theta}$ and $\tau_{\theta z}$ in addition to those shown in Fig. 2. However, even in a type of tube in which the stress distribution arose from local thermal effects and varied rapidly and erratically in both the axial and circumferential directions, continuous fringes could be observed at every level with the quartz wedge and cross polaroids. Again rotating the analyzer obscured the fringes. Thus even for this extreme case it was possible to neglect the effects of the shearing stresses and regard the observed retardation as a function of axial stress.

In the general case the maximum tension σ_m is related to the maximum relative phase retardation R_m by $\sigma_m = W\sigma_0 R_m$, where σ_0 is defined as in the preceding section and W is a dimensionless weighting factor depending on the form of the fringe pattern, that is on the variation of ψ on exit with x. As we have seen (Eq. (10)) the stress distribution assumed in shell theory gives $W=3$. The minimum value of W is unity and occurs when the axial stress is constant over the cross section. If we make the approximation that over the path of a ray in the bulb wall the stress depends only on r we can calculate the fringe patterns corresponding to different simple variations of σ_z with r. Comparing the calculated patterns with those observed we determine the range of variation of W and its dependence on obvious features of the fringe pattern. In the tubes tested at the Bell Laboratories W was found to vary from 1 to 3 and to depend mostly on the position of the observed maximum phase difference between the outer and inner surfaces of the bulb wall. For routine testing W was taken as 3 when the maximum was observed in the outer part of the bulb wall, as in Fig. 4, and as 2 for a maximum appearing in the center or inner half of the wall, as in Figs. 6 and 7. If the fringe displacement had no maximum but increased monatonically from the outer to inner surface of the wall as in Fig. 6 (bottom fringe) then W could safely be taken as unity.

Although the quartz wedge gives the most information for experimental studies, we shall see in the following section that the most practical routine method of identifying overstressed bulbs involves the use of a quarter wave plate or de Sénarment compensator. The quarter wave plate is first oriented to give minimum intensity with crossed polaroids at 45° to the vertical. The tube is then inserted with axis vertical between the polarizer and quarter wave plate and the analyzer rotated to give minimum intensity at a specified point. If $\varphi=0$ and $\omega=\pi/4$ on exit from the tube the rotation of the analyzer in radians will be equal to one-half the phase difference for the corresponding ray. By this method it is possible to obtain measurements of phase difference accurate to a few degrees. Since the deviations in ω and φ are also of the order of a few degrees it is interesting to ask whether the deviations in ω and φ are augmented or diminished in their effect on the accuracy of the measurement of phase difference.

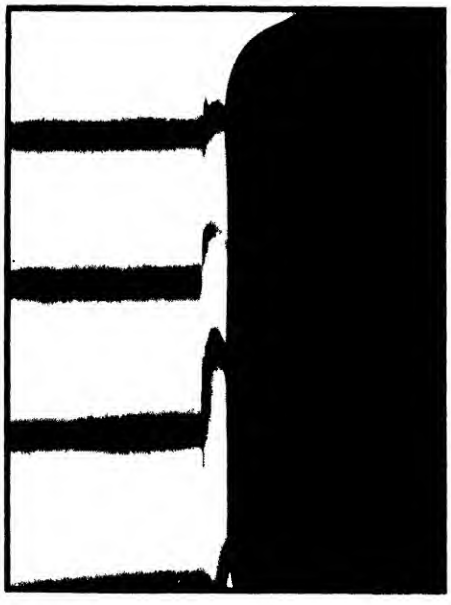


FIG. 7. Same as Fig. 6 except quartz wedge is raised to show fringes at different levels.

First we shall consider the case where $\varphi=0$ on exit and investigate the effect of taking $\omega\neq\pi/4$.

In Fig. 5 F and S are the fast and slow polarizing axes of the quarter wave plate which for $\varphi=0$ are at 45° to the f and s axes on exit. If α is the angle between the axis of the analyzer A and the S axis then it follows from simple trigonometry that the intensity of the light transmitted by the analyzer is proportional to

$$1-\cos 2[\omega-(\pi/4)]\cos(2\alpha-\psi).$$

If $\omega=\pi/4$, the light is extinguished when $2\alpha=\psi$. If $\omega\neq\pi/4$ it is impossible to obtain extinction. However, we do not set for absolute extinction but for minimum intensity which for low intensities is indistinguishable by eye from absolute extinction. It is evident from the above formula that the intensity of the light observed through the analyzer will always be a minimum when $\alpha=\frac{1}{2}\psi$ independently of the value of ω. We thus have the result that if the quarter wave plate is oriented

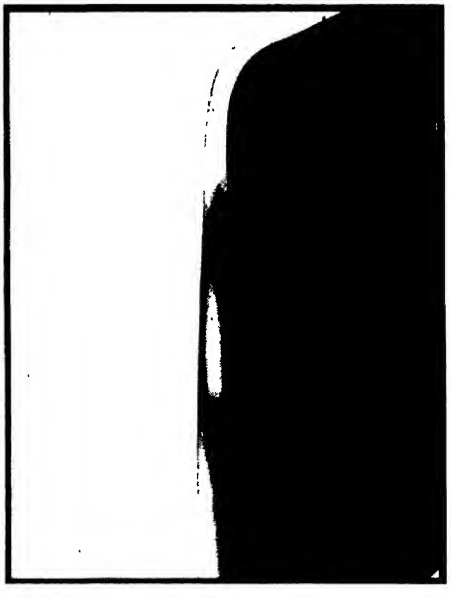


FIG. 8. Fringe pattern with quarter wave plate. Ring in bulb wall encloses region of critical stress.

with its axes at 45° to the polarizing axes of the tube on exit then the formula $\psi=2\alpha$ will be correct independently of the rotation of the polarizing axes along the path of the ray inside the tube wall.

Now consider the case where φ is not zero on exit. If we denote by ψ' the difference in phase between the H and V components we have from the preceding result that $\psi'=2\alpha$ for minimum intensity. That is, the de Sénarment compensator correctly measures the difference in phase between the H and V components. The error in the measurement of ψ is therefore $\psi'-\psi=2\alpha-\psi$. After considerable trigonometric manipulation we derive for the first-order term in $\psi'-\psi$

$$\psi'-\psi=2\sin\psi\sin\varphi\frac{\sin(\varphi+2\omega-\pi/2)}{\cos 2(\varphi+\omega-\pi/4)}$$

or neglecting second-order terms in φ and $\omega-\pi/4$ in comparison with unity we have for the upper limit of the absolute error

$$|\psi'-\psi|\leq|2\varphi^2+4\varphi(\omega-\pi/4)|.$$

Thus the error in the measurement of relative phase retardation is of second order in the deviations in ω and φ on exit and may therefore be considered negligible.

IV. TECHNIQUE OF MEASUREMENT

The de Sénarment compensator or quarter wave plate is free from a serious limitation of the quartz wedge, which gives fringes at only three or four levels in the tube wall. With the rapid and erratic variation of stress in the axial direction there is a high probability that maxima may fall between the fringes and consequently not be recorded. The de Sénarment compensator not only covers the whole extent of the tube in the axial direction but also provides the basis for a simple and practical technique for routine testing.

We have seen that when the analyzer is set at an angle α (to the S axis of the quarter wave plate—Fig. 5) a dark spot is observed in the resulting fringe pattern wherever the phase difference is 2α.

Consider the case where the analyzer is set so that 2α is equal to the maximum phase difference in the whole bulb. Then a single dark spot or fringe will be observed at the maximum and the rest of the bulb wall will

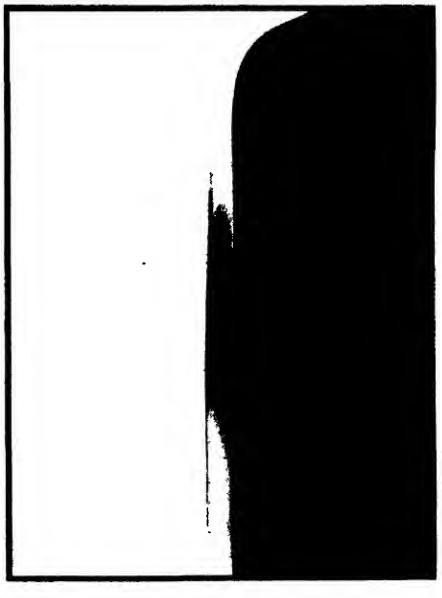


FIG. 9. Shrinking of ring about critically stressed region as analyzer is rotated.

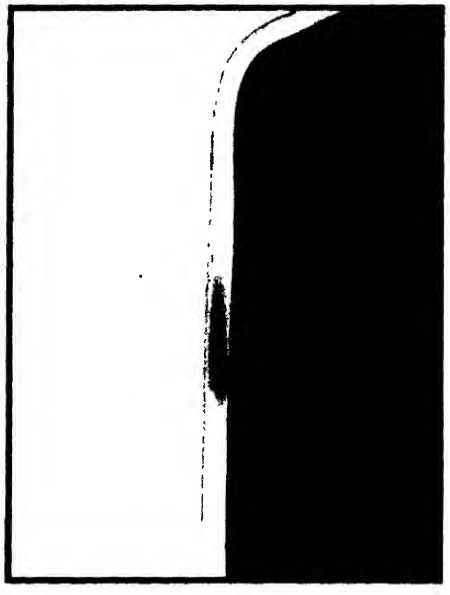


FIG. 10. Ring shrunk to single dark spot at level of maximum stress. Rotation of analyzer gives magnitude of maximum stress.

appear light as in Fig. 10. When the analyzer is set so that 2α is slightly less than the maximum phase difference the maximum will appear as a light spot enclosed by a closed dark fringe or ring on which the phase difference is 2α, Figs. 8 and 9. By rotating the analyzer so that α increases, the ring is caused to shrink to a single dark spot. The setting of the analyzer then gives the maximum phase difference.

Figures 6 and 7 are taken with a quartz wedge and illustrate a fringe pattern of typical form but somewhat larger than average amplitude. The maximum phase difference is seen to be approximately $\frac{2}{3}$ of a wavelength or $4\pi/3$ radians. This corresponds to a setting of $\alpha=120°$ for the de Sénarment compensator.

Figures 8 to 10 are taken with the de Sénarment compensator with $\alpha=60°$ in Fig. 8 $\alpha=85°$ in Fig. 9 and ending with the final setting of 120° in Fig. 10, which gives a single dark spot indicating a maximum phase retardation of 240° or $4\pi/3$ radians in agreement with the quartz wedge. The progressive shrinking of the dark ring can be clearly seen as the analyzer is rotated to approach the setting corresponding to the maximum retardation.

We now have established the following method for locating overstressed regions in the walls of glass bulbs with a de Sénarment compensator. When a limit on allowable stress has been set, the corresponding limit on phase difference is calculated (from σ_0) for the outer part of the tube wall where the weighting factor is 3. With the analyzer set for this phase difference the bulb is rotated once around its axis and the fringe pattern observed. If a dark ring appears in the outer part of the wall then the tube is rejected since the ring must enclose a maximum phase difference corresponding to a tensile stress above the allowable limit. If a dark ring appears in the center or the inside of the center of the tube wall, where the weighting factor is 2, then the analyzer setting is increased by a factor of $\frac{3}{2}$. If this causes the dark ring to shrink to a point and disappear then the tube is safe, but if a light spot still remains in the shrunken ring then the tube is rejected.

This procedure has been found to be practical for routine testing of tubes in quantity and is now in use at the Bell Telephone Laboratories at Allentown, Pennsylvania.

The experimental work involved in this investigation was carried out jointly by the author and Mr. J. E. Clark of the Electronic Apparatus Development Department of the Bell Laboratories. The photographs shown in Figs. 6 to 10 were made by R. D. Mindlin.

APPENDIX I

We shall now consider the case where both $d\varphi/dy$ and $p-q$ in Eq. (3) are constant. The exact solution for this case has been obtained by D. C. Drucker and R. D. Mindlin[5] using a dynamical theory of light.

We wish to study the dependence of the relative phase retardation ψ on the magnitude of R and the direction of polarization of the incident light. For this purpose we shall use Eqs. (12)

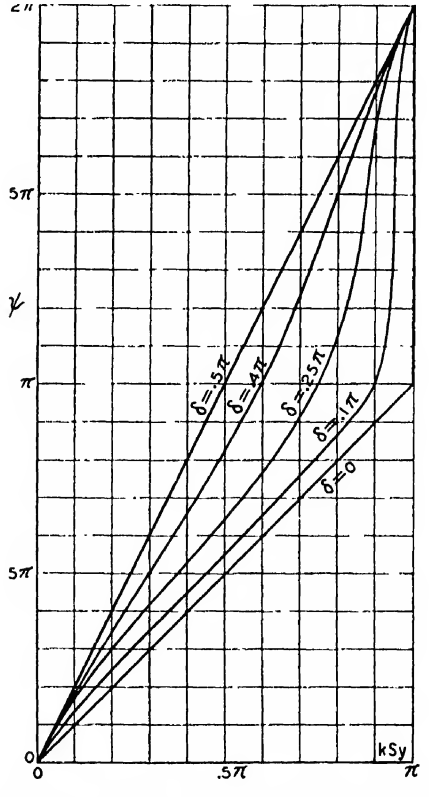


FIG. 11. Variation of phase difference with distance for several simple cases of constant principal stress difference.

[5] D. C. Drucker and R. D. Mindlin, "Stress analysis by three-dimensional photoelastic methods," J. App. Phys. 11, 724 (November, 1940).

of Drucker and Mindlin. These equations were obtained by Drucker and Mindlin upon inserting the assumption of weak birefringence in their exact solution. It can be shown that the identical formulas can be obtained by a direct solution of Neumann's equations. From these equations the difference in phase of the two waves (our ψ) is related to the distance along the path of the ray (our y) by

$$\tan\psi = \frac{\tan 2kSy}{S}\left[\frac{\tan\delta+\tan kSy}{\tan\delta+\tan 2kSy}\right], \tag{11}$$

where

$$k=\pi C/\lambda(p-q)$$
$$S=(1+R^2)^{\frac{1}{2}}$$
$$\tan\delta=\frac{S}{R\cot 2\omega_0}$$

and ω_0 is the angle between the direction of polarization of the incident light and the f axis on entry. Here we shall consider that $R^2\ll 1$ so that we may take $S=1$.

When the incident light is plane polarized at 45° to the f axis on entry, $R\cot 2\omega_0=0$ as in the simple two-dimensional case (Eq. (3)). However, as the direction of polarization of the incident light approaches the direction of one of the secondary principal axes on entry, then $\cot 2\omega_0$ becomes infinitely large, and since there is always some rotation of the secondary axes, δ approaches 0. In this case the slope of the ψ *vs.* kSy curve approaches infinity in the range $\pi\leqq\psi\leqq 2\pi$ while in the range $0\leqq\psi\leqq\pi$ the average slope is just one-half the average slope in the previous case. Figure 11 shows plots of ψ *vs.* kSy for several values of the parameter δ. In each case we take $S=1$.

Thus in this simple example where $p-q$ and R are constant and $R^2\ll 1$ the simple two-dimensional formula gives accurate results for the relative phase retardation when the incident light is polarized at 45° to the secondary principal axes on entry. However, when the direction of polarization of the incident light coincides with one of the secondary principal axes then the simple two-dimensional formula gives values of ψ which may be in error by a factor of 2.

If $S\neq 1$ then the form of the curves in Fig. 11 will be altered somewhat but the curves will still cut the lines $\psi=0$, $\psi=\pi/2$, $\psi=\pi$, $\psi=3\pi/2$, and $\psi=2\pi$ at the same values of kSy.

It should be noted that the points of equal phase will occur at intervals of π/kS independently of the direction of polarization of the incident light. If $p-q$ is calculated from measurements of the distances between points where the light is plane polarized (as in Weller's method) then the simple two-dimensional formula is correct within a factor of S (see reference 5, p. 729).

APPENDIX II

Measurement of Circumferential Stress

In this section we shall describe the method of measuring σ_θ. This method was suggested by R. D. Mindlin.

For this measurement the tube was oriented with its axis vertical, as before, but was not immersed in the liquid. Fig. 12 shows part of the tube wall bounded by sections of the outer surface A and the inner surface B. The two sets of reflected rays have approximately the same intensity since each represents the result of one reflection, and little energy is lost on refraction.

From the laws of geometrical optics it follows that the two

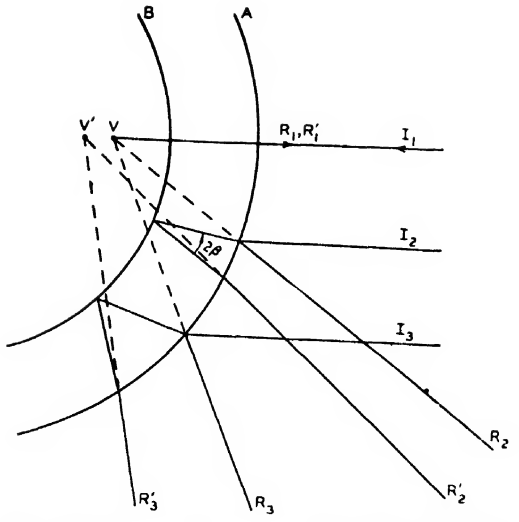


FIG. 12. Technique of measuring circumferential stress.

virtual sources V and V' are separated by a distance of $1/2n$ times the thickness of the wall where n is the index of refraction, that is the inner and outer surfaces of the tube are effectively cylindrical mirrors and the two virtual sources appear as lines parallel to the axis of the tube.

The light reflected from the outer surface of the tube remains plane polarized. However, the light reflected from the inner surface of the tube has twice passed through the tube wall and has been resolved into components which emerge out of phase.

The angle of incidence and index of refraction determine the angle β between the rays traveling inside the tube wall and the radial direction. The secondary principal stress difference in the plane of the wave front is therefore $\sigma_\theta\cos^2\beta-\sigma_z$. This gives for the phase difference ψ

$$\psi/2\pi=2\int_{-h}^{h}\frac{C}{\lambda}(\sigma_\theta\cos^2\beta-\sigma_z)\frac{dr}{\cos\beta}.$$

Since σ_z is an odd function of r it disappears in the integration and we have for the constant circumferential stress σ_θ

$$\sigma_\theta=\frac{\lambda}{4hC\cos\beta}\frac{\psi}{2\pi},$$

where as before h is the half-thickness of the tube wall. The phase difference can be measured with a quarter wave plate and analyzer inserted in the path of the reflected rays. This technique may be used for routine testing of tubes in quantity since it is only necessary to set the analyzer for the retardation corresponding to the critical circumferential stress, rotate the tube once around its axis and observe whether the virtual line source is extinguished at any point.

However, this technique cannot be used to measure the circumferential stress in any tube where the average axial stress over the thickness is not zero. However, by this technique it was shown that there was no rotation of the polarizing axes in the plane of the wave front thus indicating that the shearing stress $\tau_{\theta z}$ is negligible.

Letters to the Editor

Note on the Frictionless Bearing for Small Angular Deflections

J. A. Haringx
Philips Research Laboratories, Eindhoven, Netherlands
October 21, 1949

In the paper by Helmut Schlitt,[1] I found an interesting description of a new type of bearing for small angular deflections consisting of two concentric rings connected to each other by three radial flat springs. By choosing the radii of the rings to be equal and omitting one of the springs we obtain a construction that is known as a cross-spring pivot. In a recent paper,[2] I published the results of a calculation in respect to this constructional element and emphasized the remarkable phenomenon of its maximum rigidity for definite values of the compressive forces exerted upon the flat springs. In the paper by Schlitt I missed the occurrence of such a maximum.

In Figs. 3 and 4 of the paper[1] the curves near the vertical coordinate axis both point upward. This result is to be doubted for the following reasons. Although the separate treatment of the flat springs for tensile and compressive loads is necessary owing to the different mathematical expressions used (the hyperbolic functions change into the cyclic functions), the transition from the tensile to the compressive force is actually a gradual one, as will be evident from a physical point of view. Thus, when drawn in one single graph where the two opposite meanings of the axial load P_s are taken into account, the curves of the two separate figures for tensile and compressive loads must gradually pass into each other. Apart from the difference in the ordinates for $P_s=0$ caused by the differently chosen values of R_1/R_2, the curves may not show the sharp peak arising here.

For the case that R_1 and R_2 are equal the actual relation between the bending moment about the center and the axial force may be taken from my paper. The graph in question is here reproduced (Fig. 1); M_1 stands for M_0 and N for P_s or $-P_s$, while the other notations are the same as in Schlitt's paper. As mentioned, the rigidity M_1/φ has a maximum, at $N\approx 2\pi^2EI/L^2$. Further, in accordance with Schlitt's statement, it is seen that the rigidity can be reduced to zero by a tensile force as well as by a compressive force. The latter force is however much larger than presumed by Schlitt, namely about 80 percent of the buckling load of the spring, which for clamped ends is equal to $4\pi^2EI/L^2$. The success of the actual application of this property will therefore be questionable.

On closer examination of the calculation given it is found that some signs in Eqs. (15), (16) and (17) are mistaken and that R_1 and R_2 must be replaced by $-R_1$ and $-R_2$ respectively. Further, the right-hand member of Eq. (18) must be put equal to $-\tan(L/\lambda)$ instead of to $\tan(L/\lambda)$. This explains the observed discrepancy.

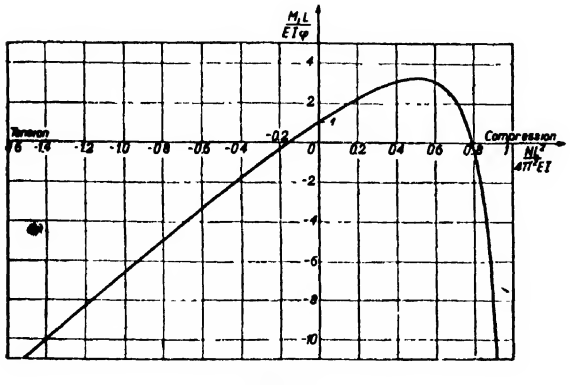


FIG. 1.

[1] H. Schlitt, "Investigations of a bearing for small angular deflections without any friction and with negligible restoring torque," J. App. Phys. 20, 297–301 (1949).
[2] J. A. Haringx, "The cross-spring pivot as a constructional element," Appl. Sci. Res. A1, 313–332 (1949).

Reply to J. A. Haringx

Helmut Schlitt
Fort Bliss, Texas
October 21, 1949

I am very thankful to Mr. Haringx for calling my attention to a discrepancy between our investigations. Concerning the physical side of the problem, I entirely agree with his considerations. There must be a gradual transition between the torque curves plotted in Figs. 3 and 4 of my paper as far as they represent bearings of equal design. Since these curves refer to different design datas of a bearing (different value of R_1/R_2), I expected the maximum of the curve to be always on the opposite side of P_s/P_{s0}. Actually there was a sign mistaken in my calculations of the compression bearing. Correcting this mistake my formulas show the changes stated by Haringx. The revised Eq. (18) yields then for the first solution for L/λ the value 5.59 corresponding exactly to the one taken from his graph.

This higher value for L/λ causes a correction of the design figures for the suggested bearing combination which is supposed to be independent of temperature. Nevertheless it seems possible to me to achieve the desired features. The fact that the compression force for zero torque of the compression bearing is rather close to the buckling load seems to me of minor importance as long as there are used three springs for each bearing. For such a design obviously no buckling can occur since the position of the two rings with respect to each other is invariably fixed.

Comments on Creep and Damping Properties of Polystyrene

William N Findley
College of Engineering, University of Illinois, Urbana, Illinois
November 1, 1949

In a recent paper,[1] Sauer, Marin, and Hsiao presented data on creep and damping capacity of polystyrene together with speculations on a correlation between damping and creep rate.

Interpretation of Creep Data: It may be of interest to consider another interpretation than the power equation used by the authors. An examination of the data given in the authors' Fig. 6[2] disclosed that the data might be represented about equally well by an exponential equation since the data formed nearly a straight line on a plot of log creep rate *vs.* stress. Because the exponential function is a good approximation of the hyperbolic sine function for relatively large values of the stress the latter function was suggested. The equation

$$v=0.26\sinh(\sigma/540) \qquad (1)$$

was found to represent the authors' data fairly well. This fact is of particular interest since there is a theoretical basis for expecting a hyperbolic sine relation which is to be found in studies of the activation process.[3]

In connection with these speculations it may be of interest to consider the creep rate data obtained by the writer[4] for tension creep tests at the same temperature and humidity used by the authors (see Fig. 1). These data were obtained from tests of a polystyrene made by the same manufacturer and having the same molecular weight of about 68,000 (Staudinger).

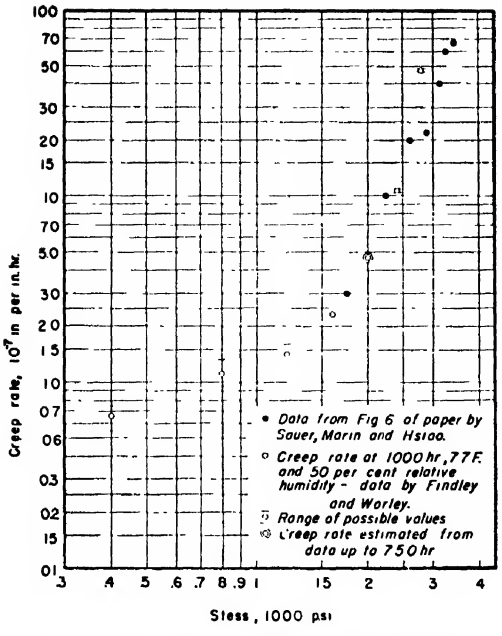


FIG. 1. Creep rate vs. stress for polystyrene in tension.

In Fig. 1 the data obtained by the authors as well as that obtained by the writer, are shown plotted on logarithmic coordinates. It is evident that the agreement is reasonably good over the interval of stress values covered in common by both tests. It is also apparent that the data obtained by the writer at the lower stresses indicate that the relationship between creep rate and stress was not a power function for the writer's tests, since the data form a curve rather than a straight line in Fig. 1.

The data obtained by the writer were represented more accurately by a hyperbolic sine function than a power equation. If the writer's datum at the largest stress is neglected, the remaining creep rates fit Eq. (1) within about $\pm 0.4\times10^{-7}$ in. per in. hr.

Interpretation of Dynamic Measurements: In view of the small scatter in the test data for damping presented by the authors, it seems appropriate to examine further the implications of these results as related to the test method and method of analysis employed, even though the method and analysis have been widely used to study the dynamic properties of many materials.

The equations used by these investigators were derived from the motion and forces present in resonant vibration of a single-degree-of-freedom system consisting of masses, linear spring, viscous damping and sinusoidal excitation. The differential equation for such a system is of the form[5]

$$M\ddot{x}+c\dot{x}+kx=F\sin\omega t, \tag{2}$$

where x, $\dot{x}$, and $\ddot{x}$ are the displacement, velocity, and acceleration respectively, M is the mass, c is the viscous damping coefficient, k is the stiffness of the spring, and $F\sin\omega t$ is the sinusoidal forcing function.

From the solution of Eq. (2) it is possible to calculate the energy dissipated per cycle, W_D, by the viscous damping.[6] The relationship obtained is

$$W_D=\pi c a^2\omega, \tag{3}$$

where a is the amplitude of vibration. For the mechanical system described above this dissipated energy is equal to the work done per cycle by the disturbing force F, as calculated by the authors, providing energy lost to the surroundings was negligible.

Equation (3) indicates that the energy dissipated per cycle by the damping is proportional to the square of the amplitude of the motion. If the stress, σ, in the specimen is understood to be defined in terms of the spring force alone and the spring force is linear as required in this solution, then the stress is proportional to the amplitude so that Eq. (3) becomes

$$W_D=C\sigma^2\omega, \tag{4}$$

where C includes the viscous damping constant and the dimensions and stiffness of the specimen.

That the tests reported by the authors showed the work done per cycle by the disturbing force (oscillator) to be proportional to the 2.3 power of stress may indicate that polystyrene behaves as a viscous material for damping. The fact that the 2.3 power was found instead of the 2.0 power may be a manifestation of the use of an equation for a system with a single viscous element, Eq. (2), to represent the behavior of a viscous material having a distribution of viscosities or relaxation times such as polystyrene is considered to have.

On the other hand it seems equally possible to the writer that the 2.3 power may result from some part of the mechanical equipment used by the authors not performing in accordance with the linear relationship required by the procedure used to calculate the damping capacity.

Further evidence that the test equipment may not function in the linear manner described by Eq. (2) are found in the facts that: (a) the so-called dynamic modulus calculated from the vibration test (5.3×10^6 p.s.i.) differs from the static modulus (4.5×10^6 p.s.i.) and in the opposite direction from that reported by previous investigators[7] using the same apparatus in tests of other plastics; (b) the dynamic modulus changed with stress. Neither of these conditions is in accord with Eq. (2) which requires the dynamic modulus to be constant and equal the static modulus.

Some of the possible sources of this disagreement may be: (A) the stiffness of the specimen changes with displacement; (B) the stiffness of the specimen changes with strain rate (or frequency); (C) the oscillator may not provide the required pure sine wave of disturbing force; (D) the damping in the specimen may not be viscous; and (E) there may be energy dissipated in non-linear friction in the grips or other connections. It is possible that the modulus increased *slightly* with increase in strain rate if the increase overlapped a portion of the spectrum of relaxation times for the material. But it seems more likely that the damping was not viscous.

An extensive study of the possible sources of extraneous energy dissipation in one type of damping testing machine has recently been reported.[8] These investigators found external losses of energy amounted to as much as 500 times the intrinsic energy dissipation of aluminum alloy specimens.

The following equation which has been considered for damping in structures[9] might prove to be more suitable in the present instance than the viscous equation for representing hysteresis damping:

$$M\ddot{x}+k(\cos b_1+i\sin b_1)x=Fe^{i\omega t}. \tag{5}$$

The damping in this equation, called complex damping, is represented in the equation by the phase angle of lag b_1.

A study of the solution of this equation shows that the damping capacity (energy absorbed per unit volume per cycle for a uniform rod vibration axially) is proportional to $\sin b_1$, the square of the stress, inversely proportional to the elastic modulus and independent of the frequency. The last three of these relationships are in accord with much of the reported data on damping capacity in materials: (A) Kimball and Lovell[10] found the damping capacity to be proportional to the square of the stress for a number of different materials. (B) They also gave rather conclusive evidence that the damping capacity is independent of frequency (or velocity). (C) The data on damping capacity reported in the literature indicates that in general materials with the highest elastic moduli have the lowest damping capacity and vice versa.

A comparison of the damping capacity described for complex damping with that predicted for viscous damping shows that the viscous damping capacity calculated from Eq. (3)

cannot be expressed (in terms of energy per unit volume) as a function of stress and factors which are properties of the unit volume of the material alone. Other quantities which depend on the particular configuration, such as the mass of the vibrating body and dimensions of the specimen tested, cannot be eliminated from the expression. Also the viscous damping capacity is not a function of the modulus of elasticity and is a function of frequency —the reverse of the relationship for complex damping.

Correlation of Damping Capacity and Creep Rate: It seems likely that damping and creep (also fatigue) are different responses to the same action at the atomic or molecular level of action in a material. Thus it would be likely that equations derived from the atomic or molecular behavior to describe these three different responses to mechanical loading would have at least one factor in common.

It would indeed be very useful if creep could be predicted from damping measurements. But to be dependable such predictions must be based on correspondence of fundamental behavior of the material under the two conditions.

The following observations may be of interest in trying to evaluate the correlation reported by the authors. For viscous behavior the damping capacity varies with the square of the stress and the creep rate varies with the first power of stress;[11] whereas the authors found a 2.3 power of stress for damping and 4.53 power of stress for creep rate.

The failure of the data for styrene to follow the same stress laws predicted for viscous behavior for both damping and creep does not seem too unreasonable in view of the differences in magnitude of strain and strain rate in the two phenomena. The behavior may result from quite different mechanisms, for example, rupture of secondary bonds between atoms in damping and movement during creep of groups of molecules of large molecular weight in a matrix of molecules of smaller molecular weight.

Dimensional considerations also suggest that the equations proposed by the authors are unlikely to be more than approximations of actual physical relations since a dimensional quantity raised to a variable or decimal power is dimensionally untenable.

[1] Sauer, Marin, and Hsiao, J. App. Phys. 20, 507 (1949).
[2] The plotted values for two points seem to differ somewhat from the values in their Table II.
[3] Dushman, Dunbar, and Huthsteiner, "Creep of metals," J. App. Phys. 15, 108–124 (1944).
[4] W. N. Findley and W. J. Worley, "Creep and fatigue of polystyrene as influenced by temperature" (not yet published).
[5] Timoshenko, *Vibration Problems in Engineering* (D. Van Nostrand Company, Inc., New York, 1937), second edition, p. 38.
[6] See reference 5, p. 45.
[7] B. J. Lazan, "Some mechanical properties of plastics and metals under sustained vibrations," Trans. Am. Soc. Mech. Eng. 65, 87 (1943). J. M. Robertson and A. Yorgiadis, "Internal friction in engineering materials," J. App. Mech. 68, A173 (1946).
[8] Cottell, Entwistle, and Thompson, "The measurement of the damping capacity of metals in torsional vibration," J. Inst. Metals, March 1918, p. 373.
[9] N. O. Myklestad, *Vibration Analysis* (McGraw-Hill Book Company, Inc., New York), second edition (in preparation), Section IV-5.
[10] A. L. Kimball and D. E. Lovell, "Internal friction in solids," Phys. Rev. 30, 947–959 (1927).
[11] Schmidt and Marlies, *Principles of High-Polymer Theory and Practice* (McGraw-Hill Book Company, Inc., New York, 1948), Chapter 7

On the Growth of Metallic Crystals from the Vapor Phase

GEORGE R. KEEPIN, JR.*
Massachusetts Institute of Technology, Cambridge, Massachusetts
November 7, 1949

GENERALLY speaking there are four methods of growing crystals: (1) from the melt; (2) from solution; (3) from metastable crystals in contact with the growing crystal in a continuous solid; and (4) from the vapor. Only the first three of these methods have been studied extensively. Recently, however, there has been increasing interest in the possibilities of the last method for growing large single crystals from certain metals and metal salts. Parameters affecting the growth of metallic crystals from their vapor phase are being studied in the hope that the results may indicate generally the conditions for growth of other metallic and metallic-halide crystals by this comparatively new technique.

TABLE I.

	Temperature of vapor gun source	Partial pressure of vapor source	Temperature of collecting surface	Pressure of nitrogen atmosphere
Cadmium	310–321°C	$\sim 10^{-1}$ mm	305±3°C	$10^{-3}-5\times10^{-3}$ mm
Zinc	409–419°C	$\sim 2\times10^{-1}$ mm	410±3°C	$10^{-3}-3\times10^{-3}$ mm

All crystallization processes may be subdivided into two main factors: the velocity of appearance of nuclei and the subsequent growth of these nuclei. The experimental conditions which have greatest effect on these two factors have been given by Sakui[1] as (1) density of vapor or temperature of vapor source, (2) temperature of collecting surface, and (3) pressure of atmosphere in which crystal growth occurs. These effects have been investigated in the case of metallic cadmium and zinc crystals grown from their vapor.

An internally heated molybdenum crucible containing the metal charge was used as a gun-type vapor source. Temperatures were measured at various points in the molybdenum gun and collector plate by copper-constantin thermocouples. The entire assembly was contained within an evacuated ten-inch bell jar. All work was done in an atmosphere of oil-pumped (dry) nitrogen which was freed of oxygen by passage through fine copper mesh at 375°C. In order to obtain the required degree of purity in the charge it was found necessary to use the spectroscopically pure metal which was then further purified by a multiple distillation procedure similar to that used by Overhage[8] in the purification of sodium metal. A silver collecting surface was found most satisfactory for cadmium while a copper or nickel surface proved best for the zinc runs. For convenience in taking data, observed crystalline forms were divided into eight categories ranging from a dense amorphous-like deposit to well-developed single crystals. Vapor gun temperatures were varied in five degree intervals from 30° below to 30° above the melting point of the metal under study. Nitrogen pressure was varied in appropriate steps from 10^{-5} to 760 mm. Collector plate temperature was increased in five degree intervals from room temperature to 10° above the melting point of each metal.

A predominance of dense amorphous-like deposit was found at low collector plate temperatures. This is to be expected since a metal vapor will condense (form growth nuclei) fastest where the difference between the vapor pressure of the solid phase and the supersaturated vapor is greatest. X-ray powder patterns of these amorphous-like deposits showed them to be microcrystals of mean diameter the order of 100A. Such studies indicated that no true amorphous deposits of either Cd or Zn were obtained in any of the work.

Conditions for the further growth of nuclei once they are formed are entirely different from those conditions necessary for their formation; thus the most perfectly developed single crystals were obtained at collector plate temperatures (T_{cp}) slightly below the respective melting points, *viz.*, optimum T_{cp} for Cd is 305±3°C and for Zn is 410±3°C. As T_{cp} is decreased from these optimum values, the amount of cadmium decreases steadily while the amount of zinc deposit, predominantly crystalline powder, was actually found to increase reaching a maximum at T_{cp}=350°C; thus we may say that the smaller zinc atoms possess a greater nuclei-forming ability than cadmium atoms in this temperature range. As T_{cp} is raised above its optimum value a rapidly diminishing yield of successively smaller crystals is obtained so that at five degrees above optimum T_{cp} values only traces of microcrystals are obtained in the case of either metal. These temperature effects are quite explicable: the mobility of atoms adsorbed by the growing crystal surface determines their rate of deposition in the crystal lattice. Accordingly, high T_{cp} means high mobility and high

rate of orientation of atoms within the lattice matrix. The potential energy of properly oriented atoms is a minimum and their chance of further migration is vanishingly small so they are effectively removed from the adsorption layer and may be replaced by new atoms arriving from the vapor gun. Orderly single crystal growth requires a dynamic balance between these two processes maintaining thereby a constant density of atoms in the adsorption layer.

For nitrogen pressures greater than 0.5 mm only traces of crystalline powder are obtained. As the pressure is decreased from 0.5 mm larger but imperfect crystalline deposits result until in the pressure interval 10^{-2} to 10^{-1} mm dendritic crystals are predominant at intermediate T_{cp} (250–350° for Cd and 200–350° for Zn). Finally, at optimum T_{cp} values well-developed hexagonal crystals were obtained in the pressure interval 10^{-2} to 5×10^{-2} mm for Cd and 10^{-2} to 3×10^{-2} for Zn. At lower pressures only small yields of dendritic, inhomogeneous crystals were found. These pressure effects become quite plausible when we recall that the surface temperature of a growing crystal is higher than that of the crystal interior due to heat of solidification and that the growth of the crystal is largely controlled by the rate of heat transfer from the growing surface. Since in a given volume cooling by convection depends critically on pressure we may expect to obtain the desired thermal equilibrium within the adsorption layer at some pressure which we here call optimum, i.e. on the crystal surface the rate of heat increase by solidification equals the rate of heat loss principally by convection.

In summary, optimum conditions have been found for the growth of large, single crystals of cadmium and zinc as shown in Table I.

Electromagnetic vibration of the collector plate at a proper frequency may represent a distinct improvement over the present method of employing thermal (random) agitation to aid atomic orientation in the growing crystal lattice. The study of other vapor crystallization processes using such a technique is contemplated.

The author is indebted to Professor D. C. Stockbarger for suggesting this problem. This research was supported in part by the Harshaw Chemical Company, Cleveland, Ohio.

* Now at Radiation Laboratory, University of California, Berkeley.
[1] S. Sakui, Tokyo Inst. Phys. Res. 34, 1131 (1938).
[2] Overhage, *Procedures in Experimental Physics* (Prentice-Hall, Inc., New York, 1945), p. 534.

Thermionic Emitting Properties of Thoria-Rhenium*

G. A. Espersen
Philips Laboratories, Inc., Irvington-on-Hudson, New York
December 12, 1949

SINTERED conducting type thoria emitters containing tungsten have been investigated by Fan.[1] Similar thoria emitters containing 250 mesh pure rhenium powder furnished by the University of Tennessee, Knoxville, Tennessee, have been studied by the writer.

Emitters consisting of 50 percent thoria—50 percent rhenium were fabricated into rods by extrusion, after which they were fired successively at 1500° C_b for 15 minutes in wet hydrogen and 2000° C_b for 15 minutes in dry hydrogen.

The emitter rods were assembled into guard ring type diodes using molybdenum anodes.

Thermionic emission data taken on a number of diodes indicated average values of $A=11.7$ amp/cm²/deg. K², $\phi=2.83$ ev for the Richardson equation.

It is interesting to note that the thermionic emission from thoria-rhenium at 1630°K is approximately 55 milliamperes while thoria-tungsten yields 0.2 milliampere/cm² at the same temperature. At 2040°K the emission for thoria-rhenium is 1.2 amp/cm² while thoria-tungsten is 0.88 amp/cm².

Activation of both the thoria-rhenium and thoria-tungsten emitters is effected by flashing them at 2400°K for approximately 30 seconds.

After 150 hours of life test at 2040°K the thoria-rhenium diodes indicated a smaller deposit of vaporized materials upon the walls of the glass bulb then the thoria-tungsten diodes.

* Partial fulfillment of ONR Contract No. N6onr-261, Task Order 1.
[1] H. Y. Fan, J. App. Phys. 20, 682 (1949).

Activation Energy for Copper Self-Diffusion

John H. Hollomon
General Electric Research Laboratories, Schenectady, New York
November 28, 1949

I HAVE read the paper by Dedrick and Gerds in the November, 1949, *Journal of Applied Physics*. From their data, plotted in their Fig. 5, they calculate 55,000 calories per gram mole—a heat of activation for self-diffusion in copper by taking the slope of the straight line they have drawn in that figure. A straight line through the points gives a heat of activation of approximately 40,000 calories per gram mole. This value is significantly smaller than any other previous experimental determinations obtained by widely accepted techniques.

Reply to "Activation Energy for Copper Self-Diffusion"

J. H. Dedrick and A. Gerds
Sylvania Electric Products, Bayside, Long Island, New York
December 19, 1949

THE authors appreciate the comment of Dr. J. H. Hollomon regarding the heat of activation for the self-diffusion of copper. It is true that if a line is drawn through all of the points in Fig. 5, the heat of activation is about 40,000 calories per gram mole as contrasted to the value of 55,000 calories per gram mole given in the paper. A recalculation by the authors shows the value to be 43,000 calories per gram mole. This is low, it is agreed, but close to the value of 46,800 calories per gram mole obtained by Raynor, Thomassen, and Rouse.[1]

However, the reason for drawing the line as shown in Fig. 5 was twofold. First, the interface radii obtained at 700°C were the smallest and, therefore, the least accurate, from the standpoint of measurement. Second, as Kuczynski[2] has pointed out, surface diffusion tends to become more effective at lower temperatures, which would effectively raise D_v. Therefore, the 700°C point was not weighted as heavily as the higher temperature points. The line as drawn thus represents the higher temperature data most rigidly.

[1] Raynor, Thomassen, and Rouse, Trans. Am. Soc. Metals 30, 313 (1942).
[2] G. C. Kuczynski, J. Metals, 169 (February, 1949).

Study of Crack Propagation Using High Speed Motion Pictures

Herbert I. Fusfeld and Josephine Carr Feder
Pitman-Dunn Laboratory, Frankford Arsenal, Philadelphia, Pennsylvania
November 30, 1949

HIGH speed motion pictures have been used to observe the process of deformation and rupture in metals at high strain rates. Preliminary studies of fracture were conducted by utilizing flat tensile specimens ½ in. wide and $\frac{1}{32}$ in. thick with grids photographically imprinted on a sensitized surface. A fixture to hold the specimens was placed in a crank press used in drawing metal sheet, and operated at a speed of 960 in./min. A Fastex camera

FIG. 1. Propagation of crack in aluminum strip under tension. Aluminum 75 SO, flat, 0.5 ×0.031 in. Time between frames = 102 μsec.

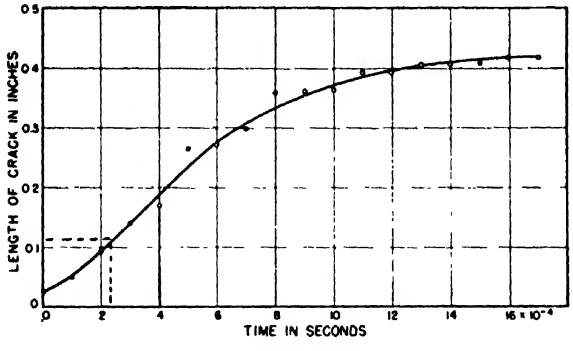


FIG. 2. Crack propagation as function of time.

was used to take motion pictures at a rate of 10,000 frames per sec.

Most of the aluminum and brass specimens tested broke along a line inclined at approximately 60° to the axis of the specimen. This break occurred too rapidly to be observed by the camera, i.e., in less than 10^{-4} sec. Thinning of the specimen along the line of fracture should be observed for approximately 5×10^{-4} sec. preceding the break.

Interesting observations have been made on the few specimens, three out of seventeen, which fractured along an irregular curve. The major portion of one such fracture is shown in Fig. 1 for aluminum 24 *ST*, showing a crack starting at the center and propagating to the edges. No noticeable thinning occurred prior to the appearance of the crack, indicating a more brittle behavior of the material than in the case where a straight line break occurs.

Measurements of the propagation of the crack are plotted in Fig. 2. The most interesting portion of the curve is at the start of propagation, as one can assume that the crack is in a uniform stress field and lines in the direction transverse to the axis of

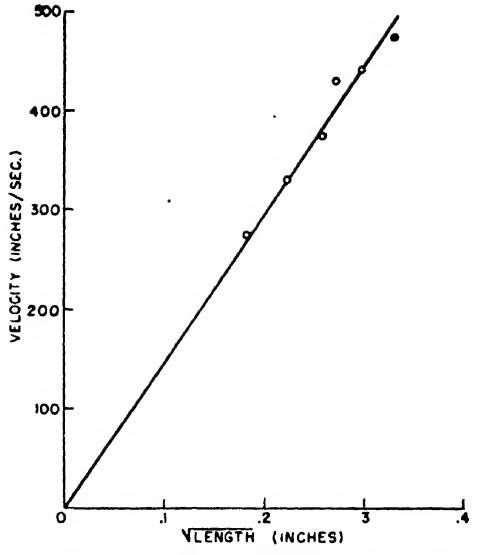


FIG. 3. Velocity of crack propagation *vs.* square root of crack length.

stress. Under these conditions, one can be guided by elasticity theory as applied to a narrow crack[1] to expect that the stress at the ends along the crack axis will increase as the crack length increases. It is therefore to be expected that the velocity of propagation would increase with the crack length, some data on which has appeared in the literature.[2]

To observe the dependence of velocity on crack length during the initial stages of propagation, the slope of the curve in Fig. 2 was measured at six points at which the ratio of crack length to specimen diameter varied from 0 to 0.64. The portion of the curve used is indicated on Fig. 2 within the dotted lines and the coordinate axes. Beyond this portion, the crack begins to extend in a direction away from that normal to the specimen axis, and the significance of "crack length" is uncertain.

A plot of velocity *vs.* length of crack on log-log paper gave a straight line with a slope of 0.47 computed on the principle of least squares. The same data is plotted in Fig. 3 as velocity *vs.* square root of crack length. A straight line through the origin provides a good fit. The slope computed by least squares is 1432 ± 26 in.$^{\frac{1}{2}}$/sec. It is therefore possible to represent the initial stages of crack propagation for the specimen of Fig. 1 by the expression:

$$v = k(s)^{\frac{1}{2}},$$

where v = velocity in in./sec.; k = slope in in.$^{\frac{1}{2}}$/sec.; s = crack length.

[1] S. Timoshenko, *Theory of Elasticity* (McGraw-Hill Book Company, Inc., New York), p. 177.
[2] G. Irwin, "Fracture dynamics," ASM Symposium on Fracturing of Metals (1948), p. 159.

Journal
of
Applied Physics

Volume 21, Number 4 April, 1950

High Speed Digital Computers

An Elementary Survey of Present Developments and Future Trends

Louis N. Ridenour

University of Illinois, Urbana, Illinois

ONE of the most interesting, and possibly one of the most significant, of the post-war technical developments is the great effort that is currently being expended, both here and abroad, on the design, construction, and improvement of high speed automatic digital computing machines. In accordance with the present pattern of support for technical enterprises, the largest part of the effort on such computers is being financed by the federal government; private industry is investing substantial sums; and, somewhat curiously, the universities of the country do not appear to be supporting with their own resources more than a tiny fraction of the work.

Mechanical aids to computation are an old story, of course. But the high speed, automatic, digital, computer—when each of these qualifying adjectives is taken properly into account—is a very new development. Its practical realization cannot be said to have been achieved earlier than the completion of the celebrated ENIAC in 1945. Of this machine we shall have more to say later. Let us pause now to distinguish from other computing devices the type of machine that will occupy our attention here.

First of all, we shall deal with *digital* machines. This excludes a large class of calculating devices customarily lumped under the name of *analog* machines.[1] An analog machine is one which translates into some physical quantity each of the numbers entering a computation; that is, the physical quantity is the measure of the number. The analog machine then manipulates these physical quantities in accordance with the nature of the computation it is desired to carry out. Finally, the physical quantity resulting from such manipulation is measured—that is, turned back into a number—to obtain the numerical answer for the computation. A slide-rule is an analog machine. Numbers are represented, on the slide-rule, by lengths which are proportional to the logarithms of such numbers. To perform a multiplication, the lengths corresponding to the two factors are added, and the resulting total length corresponds to the product, being in fact proportional to the logarithm of the product. This total length is turned back into a number by means of the scales provided on the rule, and the calculation is finished.

A digital machine also uses physical representations for numbers, of course; but instead of using a single physical quantity as the *entire* representation of the number, and thus depending upon a highly accurate presentation, manipulation, and reproduction of the magnitude of this single physical quantity, a digital machine represents by a separate physical quantity each separate digit of a number. This permits a very much greater latitude in design for the digital machine; to handle numbers of ten digits, no part of a digital machine needs to be built with a precision of one part in 10^{10} An analog machine, on the other hand, has no better over-all accuracy than that conferred on it by the precision with which it is built.

Analog machines have been brought to a high stage of perfection and usefulness. Beginning about 1931, Vannevar Bush and his co-workers at M.I.T. developed the celebrated differential analyzer, which was a mechanical device in its first embodiment. During and since the war, electrical differential analyzers of various designs have been constructed, and some of

[1] For a description of analog computers and their uses, see D. R. Hartree, *Calculating Instruments and Machines* (University of Illinois Press, Urbana, 1949), Chapters 2, 3, and 4.

these machines are currently available on the commercial market. We shall not concern ourselves here, however, with such machines. We shall speak only of *digital* computers.

In fact, we shall speak only of *automatic* digital computers. The meaning of this qualification is that we shall not deal with digital machines of the sort represented by the familiar adding machine, or the Monroe or Marchant type of desk calculator. Such machines are very old in conception and in realization, of course.[2] The first practical adding machine based on the use of number wheels was devised by Blaise Pascal in 1642; a machine capable of multiplication (by repeated addition accompanied by suitable shifts) was designed by Leibnitz in 1671 and built in 1694. In 1820 Charles Thomas, a Frenchman, designed the first commercially successful machine; with minor modifications and improvements, this machine was being built in Paris right up to the Second World War.

We shall not even concern ourselves here with the familiar punched-card machines manufactured by the International Business Machines Company. These machines stem from the work of Herman Hollerith, an employee of the U. S. Census Bureau, who in 1889 conceived the idea of using, for entering numbers into a machine and for reading out the results, holes punched in a card. A similar scheme had long been used for controlling the weaving of complicated fabrics on the Jacquard loom, and its application to computing machines had been proposed two generations earlier by Charles Babbage, of whom we say more later.

Desk calculators and IBM machines are excluded from the present discussion because such a machine performs only one elementary computation at a time, and must be instructed anew before it can perform another. That is, we can enter on a desk machine the two factors entering a multiplication, and then rely on the machine to carry out the operation and display to us the product. But, before another multiplication—or indeed any other operation—is performed, we must ourselves inform the machine what numbers enter the new computation, and which operation is to be performed on these numbers.

In an automatic machine, on the other hand, sufficient instructions are given the machine at the beginning, and the machine has sufficient capabilities within itself, to permit the whole of a complicated sequence of manipulations on many numbers to be performed by the machine without human intervention, once the signal to start has been given. Thus an automatic machine is basically different from a desk calculator; it is more nearly a mechanized simulacrum of the whole complex: human computer plus desk machine plus tables of functions for reference plus work sheet on which the course of the computation is sketched out, and intermediate results can be recorded. The idea for such an automatic machine is more than a century old. Charles Babbage, who was Lucasian Professor of Mathematics at Cambridge, conceived such a device—the Analytical Engine—and began to build it (on British government money) in 1835.[3] The Analytical Engine was never completed, largely because the mechanical engineering of Babbage's day was not equal to the task of realizing his complicated designs, but its plan was astonishingly modern in terms of the ideas which still rule the automatic computer field today.

One more qualifier is left. We shall speak here of *high speed* automatic digital computers. Earlier than the ENIAC, at least two types of automatic digital computers were designed and built. Professor Howard Aiken, of Harvard, built with the help of the IBM Company his so-called Mark I machine which was an electrically controlled, mechanically operated, automatic computer that was hailed as "Babbage's dream come true."[4] About the same time, at Bell Telephone Laboratories, a machine was built which qualified as an automatic digital computer according to the present definitions, and used electromechanical relays as elementary computing and storage elements.[5] Both these machines were of limited speed because of their dependence on mechanical elements whose motion was necessarily slow in comparison with the actions in electrical circuits. In the ENIAC,[6] on the other hand, the operations within the machine are conducted entirely in terms of the circulation through electrical lines of electrical pulses, generated at a base rate of 100,000 per second, and in terms of the registration of the effects of such pulses by means of electronic "flip-flop" circuits which can rest stably in either of two stationary states, yet be transferred from one state to the other in a fraction of a microsecond by a pulse of the correct sign.

The time required in the ENIAC to add together two numbers, each of ten decimal digits, is 200 microseconds—200 millionths of a second. In terms of the times required by machines now under construction, this is long; addition times of about 10 microseconds are common in modern designs. But, in terms of the operation of machines having mechanical elements, this is very fast indeed.

In the remarks to follow, then, whenever the word "machine" is used, it will be understood to refer to a high speed automatic digital computing machine. Let us now turn to some general considerations on the overall design of such machines.

---

[2] For a historical account of the development of desk computing machines, see the article "Calculating machines" in the *Encyclopedia Britannica*.

[3] H. P. Babbage, *Babbage's Calculating Engines* (Spon, Ltd., London, 1889).

[4] *Annals of the Harvard Computation Laboratory* (Harvard University Press, Cambridge, 1946), Vol. I.

[5] S. B. Williams, Bell Lab. Record **25**, 49 (1947).

[6] H. H. Goldstine and A. Goldstine, Math. Tables and Aids to Computation, **2**, 97 (1946); D. R. Hartree, Nature **158**, 500 (1946).

GENERAL REMARKS ON MACHINE DESIGN

While any given machine must be designed as a whole, it is often useful to distinguish as separate various parts of the machine which have various functions. The more important of such parts are:

First, the *arithmetic unit.* This device performs the individual arithmetical operations as required by the schedule of the computation being conducted. It is a very close electronic parallel to the familiar desk computing machine.

Second, the *inner* or *high speed memory unit.* This, which is presently the least satisfactory part of machines in existence or under design, is a device for registering numbers in a permanent way—either the numbers entering the problem or intermediate results—and keeping them accessible for use on demand as the computation proceeds. Any number in the memory must be available on demand in a very few microseconds. The memory is ordinarily so arranged that it can also store *orders,* which are instructions to the machine that govern the course of the computation.

Third, the *control unit.* This part of the machine is in charge of the whole operation. It keeps track of the calculation, determines which individual operation should be performed next, and causes its execution.

Finally, an *input-output unit.* This is necessary to enable the machine to communicate with its human masters. Through this unit, the machine is supplied initially with the data and instructions entering the problem; through this unit, the machine can also read out intermediate or final results. Often, the input-output unit is used to provide a large-capacity low speed storage (on magnetic wire or punched type) for numbers and orders. This outer memory supplements the inner memory, since the latter is seldom as capacious as one would desire.

When a mathematical problem is to be presented to such a machine, it must be programmed; that is, broken down into the individual steps to be performed successively. The numbers entering the problem must be supplied, and so must the *orders,* or instructions to the machine for handling those numbers. The input unit reads the program of numbers and orders into the memory unit of the machine at the beginning, and the work thereafter proceeds entirely within the machine.

A typical order might be translated: "Take the number in stored in memory location 1100 and multiply it by the number stored in memory location 2016. Put the product in memory location 1234 (erasing what stands there now) and then go to memory location 1161 for the next order." Orders are expressed in a code which gives them the appearance of numbers, so far as the machine is concerned. This has two advantages: each memory location can store either a number or an order, indifferently, and arithmetical operations can be performed on the pseudonumbers which are orders, thus changing one order into another when this is required as the computation progresses.

The example of an order just given calls for a multiplication. Similar orders govern addition, subtraction, multiplication, and sometimes certain other arithmetical operations. Another type of order, whose use greatly extends the scope of the machine, is the so-called "*conditional transfer*" order. This provides the machine with a sort of judgment, and enables the further course of a computation to be determined by the results to date, without the explicit intervention of a human operator. A typical conditional transfer order might say: "Compare the result of the last manipulation with the number stored in memory location 1648; if it is larger than that number, or equal to it, go to memory location 1174 for the next order; otherwise, go to memory location 1176 for the next order." Such decisions are a prominent feature of the control of a complicated calculation by a human computer using conventional methods.

Numbers are represented within the machine by groups of non-linear elements each of which has two stable states: on or off. For electronic machines of the present day, the non-linear elements are ordinarily the so-called "flip-flop" circuits derived from the multivibrator invented in 1919 by Eccles and Jordan.[7] Two tubes are so coupled, in a flip-flop circuit, that when one tube is conducting, the grid of the other tube is biased far beyond cut-off. If a signal is received in such a sense as to swing positive the grid of the cut-off tube, then the current in this tube rises and at the same time the grid of the other tube is driven negative. In a fraction of a microsecond, this action culminates in a stable situation in which the tube that was formerly cut off is carrying full current, while the grid of the tube that was formerly conducting is biased far beyond cut-off. Thus the system has two distinguishable stable states, and can by a pulse be transferred from one of these states to the other.

The flip-flop is not the only device which realizes practically this requirement of having two stable states which we may call on or off. We shall see later that there are other, and sometimes simpler and cheaper, ways of achieving this end. Indeed, it is in this direction that we can expect much of the short-term progress in computing-machine design to lie.

Since numbers are represented by a succession of elements which are either on or off, it seems natural to express numbers, for the purposes of the machine, in terms of the number system based on 2, rather than in the familiar decimal system based on 10. In the binary system—the base-2 system—each digit of a number has either the value 0 or 1, and the digits appearing in neighboring places with reference to the binary point (analogous to the decimal point in the base-10 system) differ in value by a factor 2, rather than by a factor 10

[7] W. H. Eccles and F. W. Jordan, Radio Review 1, 143 (1919).

as they do in the decimal system. Each number is then represented most economically, by a series of on-off elements, one for each digit of the number. The representation of decimal numbers in terms of on-off elements is somewhat less economical, since it requires the use, to represent each digit, of a collection of on-off elements so arranged that the collection has ten distinguishable stable states. These ten states are then used to represent the values, from 0 to 9, which the digit can assume.

Most of the machines presently under construction use the binary system for internal purposes; this necessitates a transformation of numbers from decimal representation to binary representation at the start and the end of each calculation, but the machine itself can perform this transformation, and the time required for this conversion is small in comparison to the time needed for the internal manipulations of numbers by a machine during the solution of a problem.

There is one other argument for the use of the binary system. Binary arithmetic is the arithmetic of logic, in a universe where a proposition is either true (one) or false (zero). The machine can, then, if it uses the binary representation of numbers, conduct its logical operations in the same formal terms as are used for its arithmetical manipulations.

Let us now pause for a moment to discuss what is called "balance" in the design of a computing machine. This term refers to the simple idea that the most efficient operation will be achieved when the times required to perform the various operations conducted by the machines are related to one another in a reasonable way. To take an absurd example, the requirement that the machine be able to perform whole sequences of computations automatically arises from its internal speed. There would be no sense in performing an addition in 200 microseconds (as the ENIAC does) or in 10 microseconds (as some of the new machines will do), if a slow-moving human operator had, at each stage, to copy off the result and then to enter the next two numbers to be summed.

At a more sophisticated level, we must consider the balance of operating times within the machine. This requires a careful weighing of the relative times required for access to the high speed memory, for addition, for transfer of a number from one place to another in the machine, and for the repeated operations involved in multiplication and division. To a certain extent, decisions on the balance of internal operating times will be made on the basis of the type of problem for which the machine will mainly be used; thus, for example, a machine whose main task was to be addition or comparison of numbers would desirably have a different internal balance of operating times from that appropriate to a machine intended for processes involving many multiplications. But the balance of a general-purpose computing machine can also be specified, and attention to this design consideration is important.

It has already been stated that numbers are ordinarily expressed in the machine in terms of their binary representation. There are two quite different ways of handling such numbers, both of which are in use in machines now building. One is called the serial method, the other the parallel method.[8] In serial operation, a number is transmitted from one place to another in the machine a single digit at a time, until the entire number is spelled out. In parallel operation, each digit of a number is transmitted on its own separate line, and all the digits of a given number are sent simultaneously.

As usual, each of these schemes has its advantages and its disadvantages. The parallel representation is inherently faster, since the entire number is transmitted at once, while in the serial scheme a number is sent one digit at a time. On the other hand, the amount of equipment needed in a parallel machine is greater than that required for a serial machine, since in the former as many lines are needed for transmitting a given number as there are digits in the number, while in serial operation all digits are sent successively down a single line. In a serial machine, exact timing is usually quite significant, since, for example, the successive digits of a number are distinguished from one another by their time of occurrence. In a parallel machine, on the other hand, timing is unimportant, since the place of a digit in a number is distinguished completely by the place at which it occurs in the machine. It is too early to say which of these schemes is the better absolutely, or the better under certain circumstances. They seem today to be competitors of approximately equal promise, and only further work and further development can determine which is the better.

The detailed design of the inner, high speed, memory has a substantial influence on the choice of serial or parallel operation for a machine. The requirements on the inner memory are very simple but very challenging. It must be able to store at least hundreds, and preferably thousands, of separate numbers or orders, and to do this in such a way that each is permanent and may be left stored for long periods without error, while at the same time it is accessible for transmission and use within a very few millionths of a second.

The earliest machines, such as the ENIAC, used for an inner memory banks of flip-flop circuits, so that at least one pair of triodes was required to store each digit of each number in the inner memory. This is such an extravagant scheme that not very many numbers could be so stored; the ENIAC has in its inner memory a capacity of only 20 numbers or orders.

It was clear that greater capacity of the inner memory is needed for achieving proper balance in a general-purpose computing machine. This problem has occupied machine designers for some time, and it still occupies them. The three types of inner memory in use are: first, a rapidly rotating magnetic drum on which the

[8] D. R. Hartree (reference 1, Section 5.6).

digits of numbers can be stored as patterns of magnetization; second, a tube containing mercury or other liquid down which the signals representing digits of numbers can be sent as pulses of sound; and third, electrostatic schemes in which the digits of numbers are stored as patterns of charge on the face of an insulator.

It is worth noticing that the magnetic drum and the sonic delay line are methods of *dynamic* storage in the sense that the numbers stored circulate continuously during storage, while the electrostatic scheme is one of *static* storage. It has therefore sometimes been said that the dynamic storage methods are best suited to a machine of serial type, while the static storage fits best a machine of parallel design. This is only partly true. A dynamic storage method produces one digit after another, to be sure; but a multiplication of the units used for dynamic storage (magnetic drums, mercury delay tubes) permits each digit of a given number or order to be entered on a separate memory unit, so that the whole number can be available at once, as it must be in a parallel machine. On the other hand, static methods of storage are not immediately suitable for use in a serial machine, but can be used if the digits of a number are read off from the static storage one after another.

Let us now turn to a consideration of the capabilities, uses, and limitations of computing machines currently under design or construction.

CAPABILITIES, USES, AND LIMITATIONS OF MACHINES

The nature of the orders presented to a machine has already been mentioned. What we must understand at this juncture is that the number of orders a machine can appreciate and act upon—its vocabulary—is built into it by its designer. An analysis of the operations of arithmetic discloses that the operations of addition, subtraction, multiplication, and division can be defined in terms of the elementary operations "join" and "meet." But the formalism necessary for this reduction is so complicated that it is more sensible to instruct the machine in terms of the four conventional operations of arithmetic than it is to reduce (in programming) every arithmetical operation to its basic terms as expressed by the operations "join" and "meet." How far can we profitably go in this direction of complicated arithmetical operation? Is it worth while, for example, to instruct the machine how to take square roots?

The ENIAC is so instructed. On a single order, the ENIAC will take the square root of a designated number. But the operation of extracting a square root is expressible in terms of the four elementary operations of arithmetic, and thus a machine which does not understand a one-word instruction to take a square root can nevertheless extract such a root if it is provided with a sequence of orders designed to produce such a result, by manipulations which the machine does understand.

This is not a trivial point. To increase the vocabulary of a machine—to increase the number of single orders which it understands—greatly complicates the design of the machine. To simplify the machine by reducing its vocabulary greatly complicates the business of programming a problem for the machine. Here, too, a balance must be struck: between machine complexity and program complexity.

In fact, we meet here one of the major differences of philosophy which has so far occurred in the design of machines. It can be illustrated by noticing that the Harvard Mark I machine is provided with a built-in table of four-place logarithms, so that it has the capability to take the logarithm of a number by looking it up in this table. The designers of other machines have asserted that, since reference to a table of function values is one of the least efficient elementary operations performed by a machine, it is better and faster to have the machine calculate for itself any function values needed, using the infinite series expressing the function. The inefficiency of tabular reference arises because either function values are stored in the inner memory (which is already too small for proper balance in most calculations, and therefore ought not to be used up for function tables), or else function values are stored in an outer memory, where consulting them takes a very long time on the extremely rapid time-scale of the machine. The storage of the form of a function makes far smaller total demands on memory space, and the computation of needed values of the function is, in general, far faster than is consultation of the outer memory to find such values. Thus, in this latter philosophy, tables of functions are an anachronism existing today only because not everyone interested in performing calculations yet has access to a high speed automatic digital computing machine.

Of course, the user of a machine often meets a situation in which he must perform repeated calculations of the same kind. If this calculation is not one comprised in the vocabulary of the machine, it must be programmed for the machine in terms of the elementary operations which are encompassed in that vocabulary. Once this is done, then the program which results can be preserved and used again and again. That is, an extended computation will ordinarily consist of a major routine and a number of sub-routines, the latter being of such a nature that they can be used, if necessary, over and over again as the main computation progresses. If, for example, the machine is not designed to understand an instruction directing it to take the square root of a number, then the square-root operation can be programmed in terms of the four basic operations of arithmetic. If that program is saved, it can be used over and over again whenever the major computation demands that a square root be taken. As a machine is used for extended periods, the number of

such sub-routines which are built up and filed away for future use will be extended; and it is not too much to expect that programming will thereby be greatly simplified.

It is of some interest to compare a high speed automatic digital computing machine to the human brain, and such comparisons have often been made in the past.[9] Such a comparison must be very carefully made. From what we know now, the central nervous system of animals, including men, is composed of individual neurons or nerve cells which are on-or-off devices like the flip-flop. However, man's central nervous system has ten thousand million such elements, while the most complicated computer so far built has only about ten thousand. This factor of a million which separates computing machines from brains is most important. For one thing, it permits the brain to perform its operations with a very considerable redundancy, so that a failure in a single element of a single chain of neurons is not important in terms of the result.

Present day machines, on the other hand, operate in a unitary way. The machine has so few "neurons" that it must use each as if it were infallible. This puts a fantastic premium on reliability in a computing machine. Von Neumann has estimated that, if a machine of the present complexity and design is not to commit errors more frequently than one per four hours, on the average, then its individual parts must have such a reliability that one failure occurs per million million elementary operations. This is some thousand times better reliability than has so far been achieved in conventional telephone practice, where one failure occurs, on the average, per thousand million operations. One of the unanswered questions about computing machines is whether such a reliability can be achieved.

This question has another side. We must ask whether such a reliability needs to be achieved. Are there not methods of checking which can, at every juncture, protect a machine from the consequences of error?

The answer is that there are. Claude Shannon, in his important paper on the "Mathematical theory of communication," has pointed the way to a method of checking which can correct automatically the errors that may be committed by a computer.[10] The scheme involved, which is due to Hamming, requires, to be sure, that other numbers beside those involved in the computation be handled and transmitted, but ensures that failures in the machine will not affect the results of a computation. In effect, the use of such a scheme introduces redundancy for the sake of reliability, and thus moves the logical design of computing machines a step nearer the logical design of the brain. The advantage of Hamming's method is that it provides a prescription for introducing the needed redundancy in the most economical and efficient way, as can be proved by Shannon's communication theory.

Probably one of the most promising directions of development now visible is this matter of evading intolerable requirements for reliability by suitably chosen means of checking. Various rather simple-minded schemes for checking have been proposed in the past—for example, to put the same problem simultaneously on two independent machines, and then to compare the results—but the checking method mentioned above is the first to have a sound theoretical basis.

Let us return to our comparison between computing machines and the brain. The individual on-off elements of a computing machine operate about a thousand times faster than do the neurons of the central nervous system. This constitutes the principal superiority of computing machines to brains; in almost every other respect the computing machines which we now can build are incomparably inferior to the nervous systems which we all possess.

We have already noticed the greatly superior complexity of the brain. Warren McCulloch, a professor of psychiatry at the College of Medicine of the University of Illinois, has observed that the complexity of the ENIAC, the most complicated machine yet built, is about that of the nervous system of the flatworm.

In terms of space and power, there is no comparison. Your brain is contained in your skull, and it dissipates less than twenty-five watts even when it is in full activity. The ENIAC, a million times less complicated, fills a large room and uses 120 kilowatts of power. Twenty kilowatts more are needed to run the blowers which keep it cool.

The great disparity between a machine and a brain must be taken into account in preparing a problem for solution on a machine. In the words of Lady Lovelace, the daughter of the poet Byron (she was speaking of Babbage's Analytical Engine): "The machine can do only what we know how to order it to perform."[11] It can obey instructions exactly, but it must have instructions or it will engage in an absurdity. This means that every possible eventuality of a complicated calculation must be visualized by the human computer when he programs the problem for the machine, and each such eventuality must be provided for in terms of explicit instructions.

Professor Hartree calls this "taking the machine's-eye view" of a computation, and he cites an instance of his own failure to do so.[12] He had failed to foresee that the argument in a trial solution might become negative, since he required finally a positive argument. In the absence of explicit instructions on this point, the machine "did its best, and this was something quite sensible according to its structure and the limited instructions which *had* been given to it, but quite different from the correct small extrapolation of those instruc-

[9] Warren S. McCulloch, Elec. Eng. 68, 492 (1949).

[10] C. E. Shannon and Warren Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949), Section 17, p. 48.

[11] H. P. Babbage (reference 3, p. 44).

[12] D. R. Hartree (reference 1, Section 7.9, p. 92).

tions which a human computor would make as a matter of course."

Up to this point we have emphasized, and perhaps overemphasized, the degree to which a computing machine is dependent upon its detailed design and on the orders it receives. The impression has probably been given that these determine entirely the operation of the machine. All this is quite true, under ordinary circumstances, but there is another way of using a machine which is worthy of remark.

I refer to the so-called "Monte Carlo" type of computation.[13] This method replaces the exact analysis of a complicated problem with a sort of statistical experimentation. That is, under properly chosen safeguards to ensure randomness, solutions of the problem at hand are secured for many special cases, and the results are examined to determine the relative probability of various over-all occurrences.

This will be clearer in terms of a specific example. The Monte Carlo technique has been applied to the problem of neutron diffusion in a scattering and absorbing medium. Neutrons of a known energy distribution fall on such a medium in a beam of known intensity; at 10 cm depth (say) what will be the neutron density and velocity distribution? At each encounter with a nucleus, a neutron has a certain chance of being captured, of being elastically scattered in one direction or another, or of being inelastically scattered with energy loss of some amount or other. The probabilities of all these events vary with the energy of the incident neutron and are different for different target nuclei (for the absorber may and usually will contain more than one nuclear type). Rather than seeking an analytical solution for this very complex problem, Ulam[13] has followed the detailed history of many individual neutrons through the absorber, at each encounter with a nucleus determining the outcome by casting dice (in effect), the dice being suitably loaded to reproduce the probability structure appropriate to the event concerned. When this has been done for very many individual neutrons, the statistics of their joint behavior presumably approaches that which would result from a full-dress analytical solution of the problem.

A high speed machine is admirably adapted to apply this method, for its speed enables it to obtain a statistically significant volume of results in a reasonably short time. However, if a machine is to do this, it must be instructed how to gamble. This is accomplished by providing to the machine, at each juncture where gambling is required, a random number whose size determines the result of the play. In fact, schemes exist for the generation by analytical processes—and therefore by and within the machine itself—of "pseudo-random" numbers which have all the necessary properties of randomness. D. H. Lehmer[14] has proposed a particularly elegant and useful scheme of this sort. The pseudorandom numbers produced according to Lehmer's prescription are easy to generate in a machine, and meet all of the standard tests for randomness. After one has made such tests, it comes as a distinct shock to find that all these numbers are divisible by 17.

There are good indications that the Monte Carlo method may be one of the best ways for solving partial differential equations on a machine. Certainly its power and scope seem destined to grow.

FUTURE TRENDS IN COMPUTING MACHINES[15]

There are three important questions which concern the future of computing machinery which I should like to discuss. The first is: "Who is likely to possess large high speed computing machines in the future?" Some workers in the computing-machine field seem quite pessimistic about the ultimate wide availability of such machines, on the grounds that they are complicated, expensive, and difficult to keep in order. This is a point of view with which I entirely disagree. I fully expect that a competent high speed computer will very soon be regarded as an important and inevitable part of the research equipment of any university having even the most modest research pretentions.

Thus, the computing machine is to be viewed not as being in the category of the large astronomical telescope, which is a pleasant but optional luxury for a university, but rather as being in the category of the electronuclear particle accelerator, which is a necessity for any university desiring to cultivate modern nuclear physics. It is a trifle surprising that so few American universities have accepted this view; in fact, only the University of Illinois is presently building a machine on its own funds for its own purposes.

Of course, the wide availability and the general acceptance of high speed computing machines will be greatly forwarded by improvements in reliability and reductions in cost. Such developments can be expected to arise from the continued developments in components and the improvement in the logical design of machines.

Let us now ask: "How large, how fast, and how complicated should a large, high speed, general-purpose computing machine be?" It seems unlikely that machines more involved than the ENIAC will ever be built. More recent machine designs are more ambitious in terms of speed of operation, in terms of the size of inner memory, and in terms of the general competence of the machine. In spite of this, such new machines have fewer tubes than the ENIAC; they use these tubes harder, so to speak. I think that the answer to the question of where to draw the line in designing a general-purpose computing machine is set entirely by con-

[13] S. Ulam, Symposium on Large-Scale Digital Calculating Machinery, Harvard University (September, 1949).

[14] D. H. Lehmer, Symposium on Large-Scale Digital Calculating Machinery, Harvard University (September, 1949).

[15] L. N. Ridenour, Symposium on Large-Scale Digital Calculating Machinery, Harvard University (September, 1949).

siderations of reliability. That is, a large general-purpose machine ought to be as big and complicated and competent as it can be made, subject to the limitation that it must not commit an error oftener than once every few hours. There is no other significant limitation on the total complexity of the device; for the machines which are now in design or construction are still quite inadequate to deal with many problems which we should like to put to them.

One simple example is contained in a remark of Hartree, quoted to me by Professor Aiken. Hartree said that the fastest computing machine available today is still too slow by a factor of 10^{10} to solve the problem of the wave equation for the copper atom. Similarly, the people who are designing the air traffic control plan for the country are proposing to put computers in key locations on the airways of the country, to handle the problems involved in air traffic control. Preliminary estimates of the scope of these problems have indicated that present-day machines will be severely taxed to deal even with the volume of air traffic that exists today.

We want, therefore, to make computing machines as large and as complicated as we can, for the most ambitious machine which we can realize today is powerless in the face of problems that we can readily pose, but not yet solve.

This brings us to my final question: "How can a computing machine be made more reliable, so that its complexity can be increased without increasing the chance of failure?" Much of the answer to this question will depend on the further development and the elaboration of the checking schemes which have already been referred to. However, the obvious way to increase the over-all reliability of present-day machines is to look toward as complete as possible an elimination from such machines of vacuum tubes and electromechanical relays. These two components are presently the major sources of failure in existing machines, partly because they are so numerous, and partly because they wear out with continued use. We need computing elements that can perform the same non-linear functions as those we now achieve with tubes or relays, but elements which are far less prone to depreciation in use.

The vacuum tube also has the disadvantage that it requires large amounts of standby power. McCulloch has remarked that, if a machine built on present principles were to have as many neurons as does the brain, it would require a skyscraper to house it, the power of Niagara to light its tubes, and the full flow of water over the Falls to keep it cool. There are now visible several developments which promise to compete with the vacuum tube for this and other purposes. First, semiconductor devices of the type of the transistor seem to have great promise. Second, magnetic devices like those reported by the Harvard group at the September symposium hold out the hope that they can serve in the place of tubes, with much greater reliability and without requiring power.[16] Finally, the pioneering work of Bowman[17] on electrochemical elements for computers offers a substantial promise in this same direction.

Probably there is no more rewarding direction for computer development than to explore the possibilities of using unconventional elements in the design of such machines. If the vacuum tube can be replaced with a cheaper, simpler, and more durable device, then the over-all competence of practicable computers can be greatly increased. Such a development will being in its train a greater availability for computers of the present sort, and a wider general use of computing machines of all kinds.

---

[16] Way Dong Woo, Symposium on Large-Scale Digital Calculating Machinery, Harvard University (September, 1949).

[17] J. R. Bowman, Symposium on Large-Scale Digital Calculating Machinery, Harvard University (September, 1949).

Contributed Original Research

A Lapping Technique: To Improve the Image Quality of Electron Microscope Lenses

F. A. Hamm
General Aniline and Film Corporation, Central Research Laboratory, Easton, Pennsylvania
(Received July 25, 1949)

One factor inherent in most electron microscopes is the magnetic asymmetry (astigmatism) in the objective lens. This is often due to imperfect machining, localized defects (impurities) in the iron, or grain anisotropies. The lapping technique described herein effectively reduces the effect of imperfect machining (misalignment of components). Critical faces and the bores through the pole piece are lapped with a series of powders (alumina) which increase in fineness until the final process is that of polishing. A set of jigs and laps (brass and hard wood) are described; their specifications and use so as to properly alter the critical pole piece surfaces are also described. Electron micrographs illustrating the value of this lapping procedure in the improvement in objective lens symmetry are included.

INTRODUCTION

DURING the past few years the electron microscope has become an important research tool. Certainly one of the reasons for this is the improvement in specimen preparation techniques. These improvements are the natural results of wider applications and the realization that artifacts must often be considered. However, because most laboratories, and especially the industrial laboratories, are usually hard pressed in applying the electron microscope to their particular urgent problems, relatively little time has been spent in improving the *performance* of the electron microscope itself. For example, each laboratory may have its own condition such as vibration, stray magnetic fields, high humidity and other accidental defects which may limit the performance of that particular instrument. With the ever increasing interest in greater resolution it has become imperative to evaluate more critically the various factors that may impair the instrument's resolution. Hillier and Ramberg[1] have very adequately evaluated both inherent and accidental instrumental defects, and critical investigators have relied heavily on the information already disclosed.

It is the purpose of this paper, first, to describe a lapping technique which has been found useful for eliminating most of the prohibitive magnetic asymmetries inherent in electron microscope objectives, and, secondly, to illustrate the practical results observed after applying this technique. Cognizance of these asymmetries is seldom taken until the investigator becomes interested in resolving structures of 50A or less (or until micrographs are taken at very high instrument magnification). When the electron microscope is used at high magnification (15,000× or greater) the aforementioned defects may be readily observed. Although at times vibration is troublesome, it was observed that the most serious factor limiting the resolution of the electron microscope in this laboratory was the *astigmatism* inherent in the *objective pole piece.* This asymmetry has been almost entirely eliminated by reducing the *mechanical inaccuracies* of an *objective pole piece* by means of a lapping technique. Both projector and objective pole pieces must be handled with care; mistakes in routine cleaning or in haphazard "polishing" techniques may be costly. It is hoped that the information disclosed herein may aid others who are interested in improving their instruments.

The contributions made by Hillier[1,2] and his coworkers should be emphasized again. It is hoped, however, that this paper will stimulate other investigators to report on similar instrumental procedures which they have found to be helpful in eliminating or minimizing conditions which may limit the resolution of their particular instrument. For example, there are undoubtedly some locations which are troubled by building vibrations.

It should be understood that asymmetries in magnetic lenses may also be due to inhomogeneities in the iron (pole pieces). Variations in the grain structure or orientation will obviously affect the distribution of the magnetic flux in the lens. These factors may be especially troublesome if they occur near critical openings. This lapping technique obviously does not eliminate these sources of trouble.

Occasionally small impurities are removed from the faces and the bores of the pole pieces. The small "pits" that remain should be eliminated insofar as possible by lapping until these surfaces are essentially smooth. These mechanical impurities are sometimes quite troublesome

[1] James Hillier and E. G. Ramberg, "The magnetic electron microscope objective: contour phenomena and the attainment of high resolving power," J. App. Phys. 18, 48–71 (January, 1947). James Hillier, "On the adjustment and manipulation of the electron microscope," Paper 32, EMSA Meeting, Toronto, Canada (September 9–11, 1949).

[2] J. Hillier and E. G. Ramberg, "Further studies on the magnetic electron microscope objective," Paper No. 41, joint meeting of EMSA and ASXRED Pittsburgh, Pennsylvania (December 7, 1946).

TABLE I.

A. Jigs (Brass)

(1) One jig to hold (with a slide fit) the sections of both pole pieces when lapping their inside faces. The broad face of the jig that makes contact with the flat glass surface is three and one-half inches in diameter. This face is counterbored 0.031 in. deep with a 1.625-in. diameter. Item 6 in Fig. 1.

(2) One jig equipped with a 0.125-in. bushing. This jig is used to hold the pole piece sections when the bores (holes through) are lapped. The bushing guides the shank of the lap. Item 7.

(3) One jig to hold the short section of the objective pole piece when lapping the outside face. This jig has a hole (0.255 in.) to guide the lap. Item 5.

B. Laps[a] (All laps are about 1 in. long; over-all length including the 0.125-in. shank is 3–4 in.)

(1) *Objective pole piece*
- (a) 0.0610 in.–brass.
- (b) 0.0615 in.–brass. Item 2.
- (c) 0.0620 to 0.0625 in.—oak wood. Item 1.
- (d) 0.255 in.—brass; brass with soft solder babbit on one end—used with *A* and *B* powders. Item 3.

(2) *Projector pole piece*
- (a) *Short section*
  - (i) 0.0630 in.–brass. A section (0.0680 in.) was left ahead of the shank which could have been turned down to the required size if necessary. Item 9.
  - (ii) 0.0650 to 0.0655 in. oak wood. Item 11.
- (b) *Long section*
  - (i) 0.0700 in.—brass. Item 8.
  - (ii) 0.0710 in.—oak wood. A 0.0720-in. portion was separated from the first (0.0710 in.) portion by means of a deep undercut. The larger diameter lap was included in case it might have become necessary. The undercut prevents the larger lap from rubbing the inside pole piece face. Item 10.

C. Powders

Exolon[b] (aluminum oxide) 600 and 800 abrasives are used to lap. The soft iron in the pole piece is removed.

Linde[c] types *A* and *B* (finest) polishing powders (aluminum oxide) are used on the wooden laps. Little or no metal is removed with the *B* powder. Minute scratches are removed probably by a plastic flow type of deformation. The particle sizes are very uniform; the *B* type is reported to be less than 0.1 micron.

D. Fluid

Singer[d] high grade sewing machine oil was used to make pastes with the powders

[a] 600 and 800 abrasives used with the brass laps, *A* and *B* polishing powders used with the oak laps.
[b] The Exolon Company, Tonawanda, New York
[c] The Linde Air Products Company, unit of Carbide and Carbon, New York, New York.
[d] Singer Oil, Type D Stainless Heavy Grade, Singer Sewing Machine Company, New York, New York.

in that more time is required to lap until a glossy surface is observed. It is conceivable that the effect of impurities might never be effectively reduced in some pole pieces; consequently the use of a compensating pole piece spacer would become imperative.

It would be difficult to evaluate the relative importance of grain structure, orientations, impurities, mechanical inaccuracies, etc. Probably these factors are all present in all pole pieces in varying degrees. However, it is worth while to eliminate the last one insofar as possible. The electron microscopist has little control over the other sources of trouble.

Because the lapping technique was done critically and with good success, it was considered worth while to report on it, giving at least the essential details. It should be emphasized that lapping the *projector* pole piece does *not* affect resolution, but it is desirable because it does remove contamination, and it may make projector lens distortion more symmetrical.

DISCUSSION

This paper is concerned only with the *practical* aspects of improving the magnetic asymmetries inherent in perhaps all objective lens pole pieces. The *theoretical* aspects are described in an earlier reference.[1] The term *astigmatism* describes a difference in focus in two mutually perpendicular directions. This difference was originally estimated to be as high as 50 microns in some objective pole pieces. This *degree of astigmatism* could have been more accurately determined by plotting objective coil current against known object displacement. Thus the difference in lens current required to bring the image in one plane into the same focus initially observed along its mutually perpendicular plane would be a measure of the degree of astigmatism as read from a predetermined curve. Such determinations are time consuming and unnecessary *in practice*. A hole in a thin membrane (collodion, silica, polystyrene, etc.) or a round particle such as carbon black (camphor) serves as a convenient subject for evaluating objective lens asymmetries. High instrument magnification is a prerequisite to a study of this kind, and it is assumed that the terms "in focus," "overfocused," and "underfocused" are understood (precluding further discussion).

Since this manuscript was first written a very pertinent investigation has been presented.[3] Through the courtesy of James Hillier the asymmetries exhibited by most of the electron micrographs have been quantitatively evaluated. These asymmetries, expressed in microns, are included in the caption for each figure. These values represent the difference in focus between mutually perpendicular directions. Very slight differences of $\frac{1}{2}\mu$ can be detected.

LAPPING TECHNIQUE

The equipment[4] used in this lapping and polishing technique is listed in Table I with a brief description. The various items are illustrated in Fig. 1.

Hillier[1] has very briefly described his procedure. This technique differs from his in three important respects: (1) The use of a 400 (coarse) lapping powder appears to be undesirable because it sometimes "tears" the soft iron. (The 600 powder removes metal at a good rate.) (2) Unless the pole pieces are dried *immediately*, water corrodes the newly formed iron surfaces. Furthermore,

[3] James Hillier and Henry Froula, "On the testing of electron microscope objectives," Paper No. 4, RCA Laboratories, Princeton, New Jersey, presented at the EMSA meeting, Washington, D. C. (Oct. 6–8, 1949).

[4] The items first used had been made with good precision by the Beeman Manufacturing Company, 1030 Haddon Avenue, Collingswood, New Jersey. Modifications and additional items have since been made in our own machine shop.

it is difficult to maintain a stable water paste of the lapping powders because of evaporation. The highly refined Singer oil does not corrode and lends itself to making stable slurries of varying consistencies. This oil is readily removed with acetone. (3) Straight laps are preferred to tapered (1°) laps because considerably more metal must be removed from one end of the bore when using a tapered lap. This invites trouble. Furthermore, the bores through the pole pieces might become abnormally large if the lapping is repeated. At least 0.004 in. of metal would have to be removed from one end of the bore, whereas only about 0.0015 in. of metal was removed with the straight lap using the 600 powder. Although no deep pits were encountered during the processing of two sets of lenses, it is conceivable that large impurities (pits) might be troublesome. If this is the case, tapered laps would be more desirable because they would preclude the necessity of making several straight laps. Obviously, the amount of metal removed by a lap is a function of the size of the particles in the lapping powder.

It would be desirable to polish with a soft metal lap such as copper or lead. However, this is not feasible because these metals are not rigid enough. Hard wood (oak, maple) was found to serve better than brass for the final polishing with the fine powders.

Before beginning the lapping it is imperative to mark both sections (of each pole piece) with respect to some mark (scratch) on the brass spacer. The pole pieces should always be assembled in the same way using these markings as a guide. This should be done before inserting the pole pieces in the electron microscope or before lapping the bores. The bores through both sections of the *objective* pole piece should be lapped with the pole piece *assembled*.[5] This is not necessary with the projector pole piece; the bore through the long section has always been larger (0.070 in.) than that through the short section (0.063 in.). Should the bores in the objective pole piece not exhibit good alignment, lapping should be continued until good alignment is realized.

A sufficient supply of a fairly thick paste of each powder is prepared so that all surfaces may be lapped before proceeding to the next finer powder. The inside faces (within the spacer when assembled) were lapped first with the 600 powder. The pole piece section does not rotate within the jig; the lapping is done with a rotary motion on a thick glass plate, preferably of good quality (flat). The assembly is lifted occasionally from the plate, turned to a new position, and then lapped as before. The pole piece is also rotated within the jig to eliminate any results due to possible imperfections in the jig. If the assembly is periodically rotated and the lapping done with a rotary motion, all directional effects will be eliminated. This same procedure is carried out with all subsequent grades of powder. All traces of the powder last used should be removed before proceeding to the next one. This can best be done by using lens tissue[6] wet with the oil, wrapped around a toothpick when cleaning the bore. A little pressure (index finger) is used in lapping the faces. The outside of the pole piece may be lubricated with the oil so that it is more readily pressed through the jig (slide fit).

After lapping the faces, especially with the 600 powder, a "burr" usually forms around the lip of the bore. This burr should *not* be removed by inserting the bore lap from either end. Apparently the lap (brass or wood) picks up the soft iron in relatively large pieces; then on withdrawing the lap the bore is scratched and sometimes badly torn. Instead, these burrs are best removed from the entire bore periphery by rolling up a small piece of crocus cloth into a cone shape, inserting the apex of the cone into the bore and gently turning the cone against the burr. At no time was there any evidence of a bevelling action on this lip. *Both* the bores and the faces are lapped to give as smooth a finish as possible with the given powder before proceeding to the next finer powder. Burrs are not usually formed when the finer powders *A* and *B* are used.

Early in this investigation it was observed that the smaller faces (long section of objective pole piece) could always be polished to a higher gloss than the larger faces. It appeared that the larger face was picking up "dirt" from the brass jig which then "stained" the pole piece face. To eliminate this, the large face of the jig, which makes contact with the glass during lapping, was counter-bored 0.031 in. deep to a diameter of 1.625 in. The pole pieces were then lapped with a rotary motion using a *short stroke* so that the lapping paste remained within the counter-bore. This procedure permitted the polishing of the large faces to the same gloss that had been observed earlier on the small faces.

The outside face of the short section of the objective

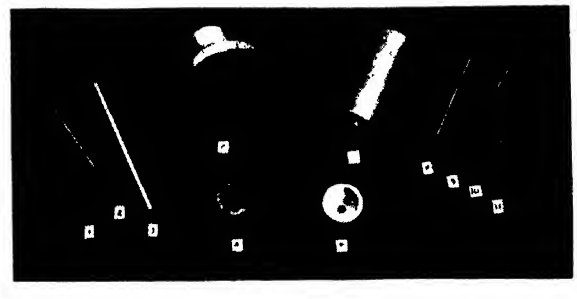


FIG. 1. Laps and jigs described in Table I.

[5] Since originally submitting this paper for publication it has been learned that some RCA e.m.u. electron microscopes have been equipped with objective pole pieces having sections with bores of unequal size. Thus, each section might be lapped separately, but it is almost imperative to lap this pole piece assembled to insure proper alignment of the openings. Consequently the bore through the one section must be carefully enlarged to the same size as the other one.

[6] Ross-Adams lens tissue sold by Clay-Adams Company, Inc., New York, does not readily tear and will not scratch the pole pieces.

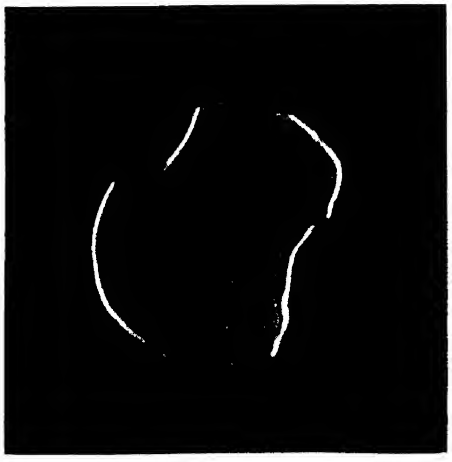


FIG. 2. 100,000×. Diatomaceous earth particle. Severe astigmatism. Note the several orders of Fresnel fringes in the vertical direction. Asymmetry of 20-25μ.

pole piece tarnishes very quickly because it is just below the specimen and because it is sprayed with electrons scattered by the specimen. This face should be polished occasionally, but care must be exercised so as to avoid introducing asymmetries. Furthermore a burr may form around the periphery of the bore when the bore is lapped; this must be removed. To accomplish these purposes, a jig was made to fit over this short section. A hole (0.255 in.) in the exact center guides a brass lap of similar dimension. This lap, used with the 600 and 800 powders, is used to lap the flat face and remove the burr. The lap must be "faced-off" (in a lathe) occasionally because the burr may cut into the brass lap. To polish this pole piece face, a similar lap is made, equipped on one end with a soft solder babbit which must also be faced-off periodically. The *A* powder embeds in the babbit so it must be removed before using the *B* powder. It is suggested that this babbit lap be used quite frequently with the *B* powder to remove the contamination that is the result of the effect of the electron beam.

Lapping the bores of the pole pieces must be done with caution. The hole must be enlarged gradually and symmetrically. The early work on the bores suggested the use of the 600 powder instead of the 400 powder. The Beeman[4] jig served its purpose very well. The 0.125-in. bushing (1.5 in. long) guides the shank of the lap so that it runs "true." The hole for the pole piece is about $1\frac{1}{2}$ in. long (slide fit) so that if the laps are properly made the finished bore should be perfectly symmetrical. In practice the abrasive paste is placed within the bore before inserting the pole piece in the jig. The clean lap is then inserted through the bushing (lubricated) and lapping carefully started when the lap meets the bore. The ends were pointed, merely to aid in starting the lap in the bore. The objective pole piece was inserted in the jig so that the lap went through the short section first.

Lapping is accomplished by rotating the lap back and forth through an arc of about 180°C. The lap is periodically withdrawn, rotated slightly, reinserted, and lapping continued as before. This ensures symmetrical action around the entire bore. The pole piece is tightly held in the jig with the index finger. Occasionally large scratches or tears developed during the early work. This was very probably caused by some hard grit which had accidentally gotten into the system.

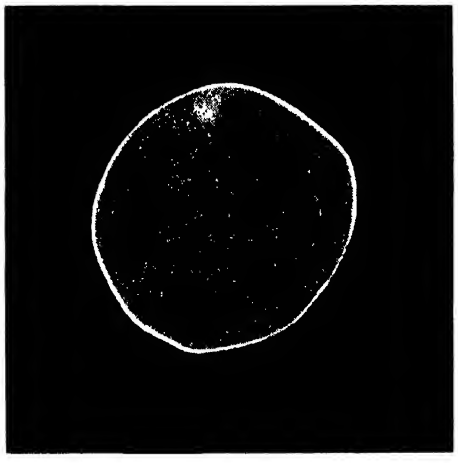


FIG. 3. 112,500×. Compare with Fig. 2. Very slightly underfocused. Different lens. 4-5μ.

The later work was carried out in a cleaner environment, and no trouble was experienced. The bore diameters of the pole pieces in the RCA e.m.u instrument have been varied slightly during the last few years; the bore in the long section of the projector pole piece is apparently always the largest. These bore diameters should, of course, be carefully measured before making the laps.

The hard wooden laps used for the final polishing show signs of wear after some use. By cutting a few grooves (with a razor blade) parallel to the long direction of the lap and soaking the lap in oil, a sufficient increase in diameter results so that these laps may be used repeatedly. It might be expected that these wooden laps do not wear evenly. Although this may be true at times, it is of no significance when polishing with the fine powders.

These wooden laps were used very successfully for polishing the bores. For this reason hard wood laps were made to polish several other surfaces. For example, the bores through the specimen holder and the special Armco iron sleeve in the intermediate screen assembly are occasionally polished with wooden laps (oak) to remove tarnish. This procedure is desirable because polishing these surfaces makes them smoother so that they are more readily cleaned with lens tissue in the routine way employed by many investigators. In both cases, the wooden laps may be held in a small drill press; the polishing is then done by moving the metal

surface up and down over the lap. Hand polishing is preferred because better control is realized.

The cleanliness of the inside conical face of the short section of the projector pole piece is very important, especially at high instrument magnification. At high projector lens current values the imaging beam of electrons is spread so that many of the rays strike the inside of the bottom (short) section of the pole piece. The final image may be of a poor quality if the beam is affected by contamination which may acquire a charge. A wooden lap (maple) was made to polish this face with the *A* or *B* powders. This lap is illustrated as item 4 in Fig. 1. The conical shaped portion (45° cone) is 0.853 in. long, with the sharp apex of the cone rounded. The short section of the projector pole piece is screwed into the brass spacer. The spacer is fitted into the jig

TABLE II. Instrumental operating conditions.

(1) Self-biased gun (saturating between 200 and 250 micro amperes), filament distance from grid (gun focusing aperture) about 4 mm. "Home made" filament sharp apex bent over a razor blade. "Cross-over" image of source approximately normal to the astigmatic plane.
(2) Primary fluorescent viewing screen removed. Brass plate substituted for the usual primary viewing window and copper screen. These changes eliminate some contamination, which may subsequently charge up and cause movement of the beam.
(3) Condenser current[a] = 84 milliamps; this provides an angular aperture of illumination to the specimen of about 4×10^{-4} radian.
(4) Mu-metal shielding between objective and projector lenses to eliminate stray magnetic fields.
(5) Intermediate fluorescent viewing screen removed. Armco iron sleeve in intermediate screen assembly to maintain magnetic symmetry within the upper part of the projector lens (see reference 1).
(6) Objective lens focusing circuit modified to give threefold sensitivity over the standard e.m.u.-2A circuit.
(7) Strict cleanliness of entire microscope column (gun) observed.
(8) Voltage alignment always used. Magnetic alignment is not superimposed; magnetic center of rotation just off the final viewing screen.
(9) No objective aperture.
(10) The micrographs were taken at tap 9 or 10 using the high gain projector spacer.

[a] Some of the data were calculated from information received in a private communication with Dr. James Hillier, RCA Laboratories, Princeton, New Jersey.

with the 0.125-in. bushing, having used tissue paper as a shim. The jig is then placed in the collet of a lathe, which is run at about 400 r.p.m. Polishing is accomplished by moving the wooden lap in and out (by hand) along one side of the cone face. A shoulder at the base of the cone prevents the cone from being inserted too far. Thus, this lap does not touch the bore. This procedure removes the surface contamination, and the tin plating, so that the copper plating is exposed. Sometimes the tin plating on a new pole piece separates from the iron *within* the bore of the pole piece. A burr of this type could be troublesome and might readily tend to catch small contaminating particles. The lapping technique described earlier effectively removes this tin plating.

After carrying out any of the aforementioned steps in the lapping procedure, the oily paste is best removed by wiping the surfaces with a good quality lens tissue.[6] After the lapping has been completed, the oil and last traces of powder are best removed by immersing the lens components in two successive baths of high grade (dirt-free) acetone. The critical surfaces should not be allowed to dry for the last time without first wiping them with lens tissue wet with fresh acetone. No evidence to suggest the incomplete removal of oil has ever been observed. Complete removal of the oil is absolutely imperative for several reasons, the most important being the fact that the electron beam would probably act on the oil (radiation-chemical) so as to cause severe contamination.

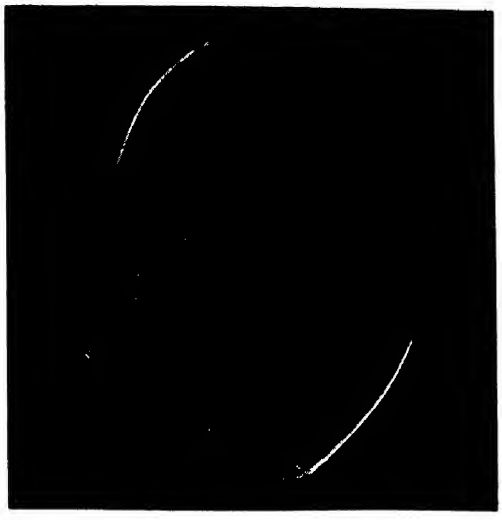


FIG. 4. 112,500×. Portion of a diatom. The astigmatism is manifest as a ghost image due to overfocusing along one direction Before lapping. Abnormally bad asymmetry of about 60μ.

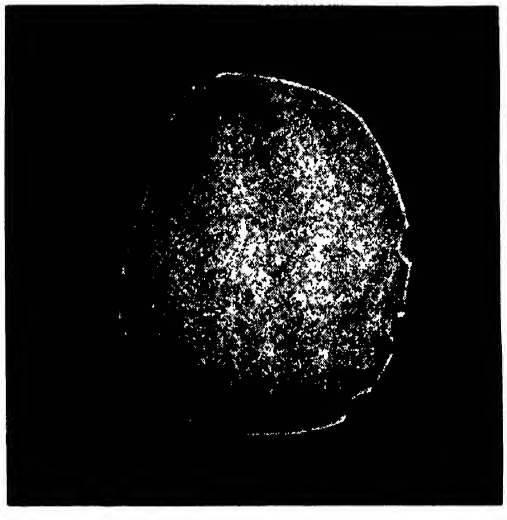


FIG. 5. 112,500×. Diatom very slightly underfocused. Same lens after lapping. 1½μ. This very small asymmetry would have been more obvious at exact focus.

FIG. 6. 112,500×. Thin film of Formvar. 1½μ. Possibly a trace of specimen drift.

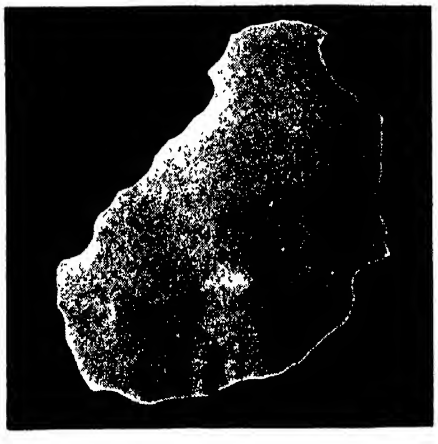


FIG. 7. 48,000×. 200A (calc.) thick film of silica. Too far underfocused to detect the 1½μ-asymmetry.

DISCUSSION OF ELECTRON MICROGRAPHS

The practical use to which the electron microscope is put is very largely based on the photographic record of the images observed. Because it was the purpose of this investigation to increase the "instrumental resolving power" most of the micrographs will be interpreted with this in mind. The "Report of the electron microscope society of America's Committee on Resolution"[7] has been used as a guide in choosing the terminology. Thus, the "instrumental resolving power" is based on measurements made on the photographic record of an image taken under ideal conditions (specimen, photographic, instrumental, etc.). The "quality of the micrograph" is more closely associated with the information revealed in relation to what was hoped for in any given investigation; although it is a function of the performance of the microscope it is probably most closely related to the quality of the electron microscope specimen. A few electron micrographs are illustrated to demonstrate their high quality as a result of better instrument performance. For the sake of convenience the "minimum distance between adjacent particle centers" is used to evaluate the instrumental resolving power. Reliance on micrographs is, of course, necessary. Most of the micrographs are illustrated in an effort to demonstrate the effectiveness of the pole piece lapping technique in eliminating astigmatism rather than to claim some specific optimum resolving power.

In order to make a fair appraisal of the lapping technique it is imperative to eliminate the effects of all other variables. Unless otherwise stated, the operating conditions apropos of these micrographs are briefly summarized in Table II.

Figure 2 illustrates an extreme case of astigmatism. The underfocused portion of the image shows at least

FIG. 8. 160,000×. Sublimed organic pigment. Poor image quality. Before lapping.

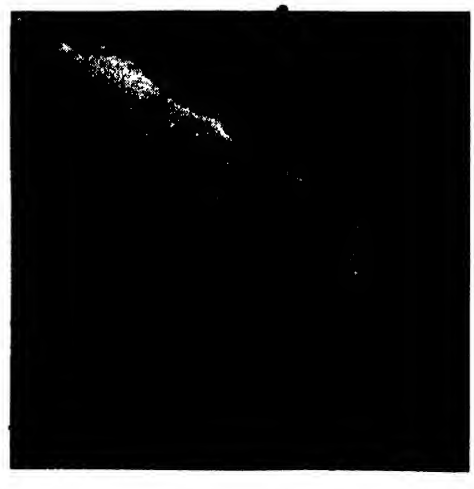


FIG. 9. 160,000×. Unshadowed silica replica of a silver bromide crystal. Good image quality. After lapping.

[7] J. App. Phys. 17, 989–996 (1946)

four orders of Fresnel fringes, indicative of good instrument performance. On the other hand, the image along the horizontal plane is in better focus (exhibiting no diffraction fringes) thereby demonstrating the astigmatism of this objective pole piece. Experience suggests that this amount of asymmetry (20–25μ) is not uncommon in many objective lenses.

Figure 3 is a very slightly underfocused image of the same specimen photographed with a different objective pole piece. The image is too close to being in focus so that only the first-order Fresnel fringe is exhibited. The symmetrical distribution of the white halo around the entire periphery of the hole (diatom) indicates a lens of good quality.

Astigmatism always becomes more pronounced as the image is brought toward exact focus. This is significant because in any critical investigation requiring high resolution, reliance on well-focused images should be made. In cases where the specimen is extremely thin and has little scattering power, a slightly underfocused image reveals more information because of its greater contrast due to phase delay phenomena. By varying the current in the vernier (fourth potentiometer) of the modified objective focusing circuit, the astigmatism can be detected and measured qualitatively in terms of vernier divisions.

Figures 4 and 5 again illustrate the effect of the lapping technique. The holes in diatoms conveniently serve this purpose. The image (Fig. 4) in the plane perpendicular to the arrow is badly overfocused (focal length too short); the parallel direction is in better focus. Figure 5 illustrates a high quality image of the same specimen after lapping. Because diatoms are relatively thick, a thin white halo appears just within the holes even though the image is, for all intents and purposes, in focus. (As a result of the quantitative evaluation[3] this image is now considered to be one or two microns underfocused.)

About one month after the pole pieces had been lapped and polished for the first time it became necessary to polish them again. The bores through and faces (especially on the objective pole piece) had become badly "tarnished." (This tarnish may have been a contamination coming from the particular type of specimen during exposure to the electron beam.) It was decided to lap the objective pole piece again lightly, starting with the 800 powder. The micrographs in Figs. 6 and 7 were taken subsequently. Although the contrast in these figures is poor (no objective aperture) the images still exhibit about the same amount of asymmetry. Because the asymmetries were almost entirely eliminated, it has been unnecessary to use a compensating objective pole piece spacer described earlier by Hillier and Ramberg.[1] It is felt that it is more desirable to eliminate the asymmetries insofar as possible by lapping rather than by using a compensating spacer because laps and jigs must be made anyway in order to occasionally remove the pole piece tarnish by *polishing*. If this is done, only a little more effort is required to complete the job by carrying out the entire lapping procedure. It is conceivable that a compensating objective pole piece spacer might be very desirable for routine work even after lapping, depending upon how badly asymmetrical the pole piece might have been initially due to reasons discussed earlier. In any case, the last traces of astigmatism can perhaps only be removed by compensation.

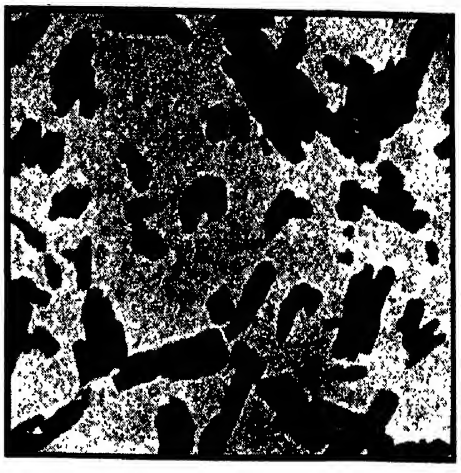


FIG. 10. 56,000×. Dispersed pigment. Poor image quality.

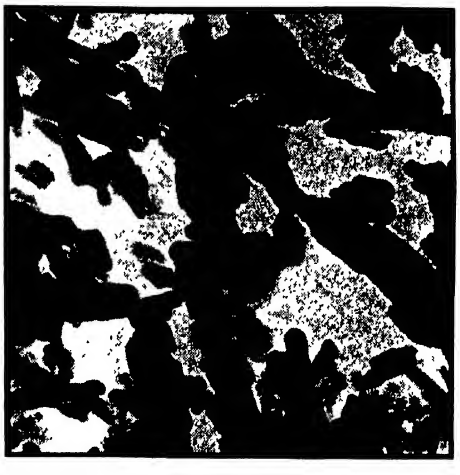


FIG. 11. 56,000×. Same preparation as above. Good image quality.

Admittedly, it is not entirely fair to compare the image quality of micrographs taken of different specimens. However, Figs. 8 and 9 are illustrated for comparison. Both types of samples are very stable in the electron beam, and both micrographs (photographed at 20,000×) represent the optimum quality possible at the time that the micrograph was taken. The difference in image quality is undoubtedly due to the difference in the performances of the lenses. The magnification in both these figures is 200,000 diameters. The absence of

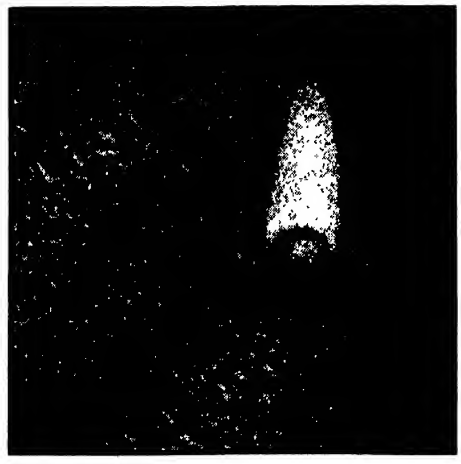


FIG. 12. 160,000×. A silicon monoxide replica of mica shadowed with Pt-Pd. Good image quality.

specimen drift, always a serious problem, should be noted.

Experience during the past several years suggests that image quality is often attributable to the performances of the electron microscope lenses. Figure 8 is several years old, whereas Fig. 9 was taken since the lenses were first lapped and polished. Note how much more "crisp" the latter image is. The contrast is poor; the specimen was not shadowed. The latter image is however, capable of resolving structure at least on the order of 25A.

Figures 10 and 11 represent a comparison of the same specimen, Indanthrene Blue RS, dispersed ultrasonically, demonstrating the "before and after" version of the lapping technique. Figure 10 is slightly underfocused. Underfocusing invariably causes the image to be more crisp and to have more aesthetic appeal. Nevertheless, Fig. 11, taken after the pole pieces had been processed, is much more crisp and suggests, even at this low magnification, the improvement in resolving power due to the lapping technique.

The structure in Fig. 12 was exhibited by a silicon monoxide replica of mica, shadowed at 10° with a platinum-palladium alloy. This structure appears as an undesirable "background grain" when the replica is used as a substrate for other objects (viruses, bacteria, etc.). The structure is typical of platinum-palladium shadowing metal, especially when the metal is relatively thick. Although the image is slightly underfocused, the quality is good and superior to that possible with an "average" objective lens. It should be emphasized that this micrograph and all the others were taken at very high instrument magnification.

CONCLUSION

The effect of astigmatism, an objective lens asymmetry, has been illustrated. A procedure for minimizing this inherent defect is outlined. This process involves lapping and polishing with fine abrasives all critical surfaces on the objective (most important) and projector pole pieces. Routine polishing with the *B* powder should be done frequently to remove contamination. High quality high magnification electron micrographs necessitate routine polishing at least once a month and preferably more often.

Improvement in the instrumental resolving power based on measurements made on micrographs was illustrated. No specific optimum resolving power is claimed. The elimination of astigmatism and the general improvement in image crispness unquestionably permit greater resolution. A reduction in objective lens asymmetry from about 60μ to about 1½μ has been illustrated.

Since this initial investigation was completed, another objective pole piece was processed in the same way. The gratifying results add weight to the significance of this technique.

ACKNOWLEDGMENTS

The author is indebted to A. McNevin for his suggestions and care in carrying out much of the lapping procedure. The interest and constructive criticisms expressed by M. C. Banca and James Hillier of the Radio Corporation of America have also been appreciated.

Separation of Gases by Fractional Permeation through Membranes

SOL WELLER* AND WALDO A. STEINER†
Research and Development Branch, Bureau of Mines, Pittsburgh, Pennsylvania
(Received September 19, 1949)

The permeability of a number of thin organic films toward oxygen and nitrogen has been measured. For an ethyl cellulose film the studies were extended to include CO_2, A, He, and H_2. The enrichment of a binary gas mixture in a single stage of permeation has been calculated for the extreme cases of perfect mixing and no mixing. Application of these results indicates that the use of a fractional permeation process may be of practical importance in effecting the separation of oxygen from air, helium from natural gas, and hydrogen from coke-oven gas, as examples.

INTRODUCTION

THE fact that non-porous rubber membranes show a higher permeability to O_2 than to N_2 has been known for almost 120 years[1] and has formed the basis for several old patents for the enrichment of air.[2] Apparently, little attention has been paid in the interim to the possible practical importance of this selective permeability behavior. Most of the work which has been reported on the phenomenon has been done on rubber membranes,[3] although scattered investigations have been made on other materials.[4,5]

In 1942, von Elbe,[6] working with stretched films of natural rubber, found that air, diffusing from a pressure of 4 atmos. to atmospheric pressure, was enriched to an O_2 content of 33 percent, corresponding to a value of 2.4 for the ratio of the permeability constants of O_2 and N_2. Von Elbe's stretched films, which had a thickness of only 0.00038 cm, possessed a permeability coefficient toward O_2 of 1.6×10^{-6} mole O_2 atmos.$^{-1}$ cm^{-2} min.$^{-1}$. Calculations based on these results indicated that a multistage permeation process for obtaining O_2 from air was economically unfavorable compared with the Linde process by a factor of about 2, with the possibility of major improvements if films of superior properties could be found. It was realized that, for possible practical use in effecting the separation of gases, films should have the following qualities: (1) high absolute permeability, to minimize the area of film required; (2) high selective permeability toward the desired gas, to reduce the power required for the separation; and (3) chemical and physical stability. Consideration of literature data[4,5] showed that many types of films probably would not be useful because of failure to satisfy the first requirement; among these were Cellophane, cellulose acetate, Nylon, polyvinyl chloride, polyvinylidene chloride, and rubber hydrochloride. It was considered desirable to test at least rubber, ethyl cellulose, polyethylene, and polystyrene.

EXPERIMENTAL PROCEDURE

The apparatus used to measure permeability was quite simple. A small Seitz bacterial filter was modified to hold the sample of film which was supported on one side by filter paper backed by 30-mesh wire screen. The exposed area of film was 18 sq. cm. In determining permeability to pure gases, the film holder, which was immersed in a thermostated water bath, was simultaneously evacuated on both sides of the film with a rotary oil pump. The pump was then closed off, the pure, dried gas was admitted on the unsupported side of the film to the desired pressure, and the rate of gas permeation was followed by measuring the rate of pressure increase in an enclosed volume on the supported (low pressure) side of the film with a tilting McLeod gauge (0–5 mm range). The enclosed volume on the low pressure side was calibrated so that the permeability could be expressed in standard units. Typical pressure-time curves, in this case for O_2 and N_2 through a 0.023 mm polyethylene film, are shown in Fig. 1.

The determination of pressure-time curves was convenient for the study of pure gases but not for gas mixtures. In several cases when gas mixtures were used, the composition of the permeated gas was determined by connecting the low pressure side of the film holder

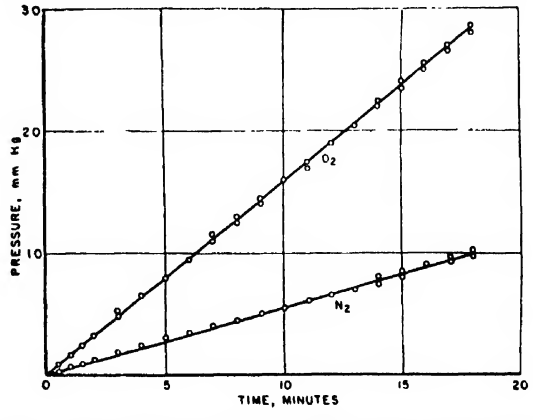


FIG. 1. Permeability of 0.023-mm polyethylene to O_2 and N_2.

* Physical chemist.
† Physicist.
[1] J. V. Mitchell, J. Roy. Inst. **2**, 101, 307 (1831).
[2] P. Margis, D.R.P.17981, Class 12 (August 7, 1881); M. Herzog, U.S. 307,041 (October 21, 1884).
[3] See, for example, G. J. van Amerongen, J. App. Phys. **17**, 972 (1946).
[4] R. M. Barrer, *Diffusion in and through Solids* (Cambridge University Press, London, 1941), Chapter IX.
[5] D. W. Davis, Mod. Packaging **19**, 145 (1946).
[6] G. von Elbe, U. S. Bureau of Mines (unpublished work).

TABLE I. Film permeability to oxygen and nitrogen at 30°C.

Material	Supplier[a]	Film thickness (mm)	Specific permeability, K, at 30°C (cm³ min.⁻¹ atmos.⁻¹)×10⁶ K_{O_2}	K_{N_2}	$\frac{K_{O_2}}{K_{N_2}}=\alpha$	Literature[c] K_{O_2} at 21–22°C (cm³ min.⁻¹ atmos.⁻¹×10⁶)
Ethyl cellulose	A	0.026[b]	4.8	2.0	2.4	3.8
Ethyl cellulose	B	0.025[b]	4.6	1.4	3.3	
Ethyl cellulose	C	0.0075[b]	9.4	3.0	3.1	
Polyvinyl chloride-acetate	A	0.051	1.8	1.2	1.5	
Vinyl resin and Buna N rubber	A	0.051	0.70	0.58	1.2	
Natural rubber latex	A	0.19[b]	18.1	8.6	2.1	11.0
Natural rubber latex	G	0.051[b]	11.4	4.4	2.6	
Resproid 1404	E	0.11	3.0	1.1	2.8	
Polyethylene	D	0.038	2.1	0.79	2.7	0.6
Polyethylene	F	0.023	2.2	0.88	2.5	
Polystyrene	D	0.025	1.3	0.44	3.0	0.8
Cellulose acetate	B	0.028	2.1	1.2	1.8	0.4
Cellulose propionate	B	0.023	2.8	0.9	3.2	
Kel-F	H	0.051	2.34	1.17	2.0	
Koroseal	I	0.038	0.37	0.15	2.4	

[a] Suppliers: A, Reynolds Metal Company; B, Celanese Corporation of America; C, cast in laboratory, D, Bakelite Corporation; E, Respro Inc.; F, Plax Corporation; G, cast in laboratory, rubber latex supplied by American Anode Inc.; H, M. W. Kellogg Company; I, B. F. Goodrich Company.
[b] Calculated; density of ethyl cellulose taken as 1.14 g/cu. cm. of rubber as 1.07 g/cu. cm.
[c] Reference 5.

directly to the intake of a mass spectrometer. The procedure was otherwise identical with that outlined. In some experiments with air, superatmospheric pressure (4 atmos.) was used, which resulted in the production of enough sample in 16 hours to permit analysis of the permeated gas by Orsat apparatus.

Most of the films studied were commercial films supplied by the manufacturer (see Table I). Films of ethyl cellulose and of rubber were also cast from solution in the laboratory with the use of Bird film-applicators. It was found possible to make a multiply drawn film of ethyl cellulose, having a total thickness of only 0.0075 mm (0.0003 in.), which did not show any leakage; i.e., the film maintained its selective permeability toward oxygen.

CALCULATION OF ENRICHMENT IN SINGLE STAGE

We assume the permeation of components A and B in a binary gas mixture to obey Fick's law, the rates of permeation of each component through the film area, $d\sum$, being

$$-dn_A=Q_A d\sum[P(n_A)/(n_A+n_B)-p(n_A')/(n_A'+n_B')]$$
$$-dn_B=Q_B d\sum[P(n_B)/(n_A+n_B)-p(n_B')/(n_A'+n_B')]. \quad (1)$$

P is the high pressure, p the low pressure, n_A and n_B the moles of A and B, respectively, flowing per unit time on the high pressure side, n_A' and n_B' the corresponding moles flowing on the low pressure side, and Q_A and Q_B the permeability coefficients (uncorrected for thickness) of A and B.

In the limiting case, where the composition of gas on the high pressure side is held equal to the inlet composition and the permeated gas is immediately removed by evacuation of the low pressure side, maximum enrichment of A is obtained (for $Q_A>Q_B$) and the ratio of molar flow rates for the components of the permeated gas is given by

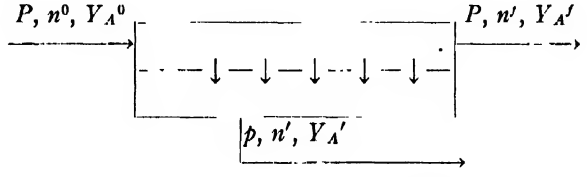
$$n_A'/n_B'=dn_A/dn_B=Q_A/Q_B\cdot n_A^0/n_B^0=\alpha n_A^0/n_B^0, \quad (2)$$

where $\alpha=Q_A/Q_B$ and n_A^0 and n_B^0 are the entering molar flow rates of A and B.

If a finite fraction of gas entering a stage permeates during that stage, the composition of the permeated gas depends on the nature of the flow conditions on the high and low pressure sides of the film. Only two extreme cases, corresponding to perfect mixing and to no mixing, will be considered.

Case I

Perfect mixing of gases occurs on both sides of the film. The conditions are represented by the following diagram.



n^0 moles of gas per unit time ($n^0=n_A^0+n_B^0$) of composition Y_A^0 [$Y_A^0=n_A^0/(n_A^0+n_B^0)$] enter the stage at pressure P. Leaving the stage per unit time, still at pressure P, are n' moles of unpermeated gas having a mole fraction of A equal to Y_A'. The n' moles of gas permeating per unit time at pressure p have the composition described by Y_A'. It is assumed that the rate of mixing on the high pressure side is so rapid, compared with the rates of flow, that the unpermeated gas has, at all points in the stage, the same composition as the unpermeated gas leaving the stage. In the steady state, the following equations of continuity and conservation must be satisfied:

$$n'Y_A'=Q_A\sum[PY_A'-pY_A'] \quad (3)$$
$$n'(1-Y_A')=Q_B\sum[P(1-Y_A')-p(1-Y_A')] \quad (4)$$
$$n^0=n'+n' \quad (5)$$
$$n^0Y_A^0=n'Y_A'+n'Y_A'. \quad (6)$$

Let $\alpha=Q_A/Q_B$, $\beta=n'/\sum Q_A$. From Eqs. (3) and (4) is obtained

$$Y_A'=\frac{\alpha-(P-p)/\beta}{\alpha-1}, \quad Y_A'=\frac{(\beta+p)[\alpha-(P-p)/\beta]}{P(\alpha-1)}$$
$$=(\beta+p)Y_A'/P. \quad (7)$$

Y^0 is assumed known, and in any particular case a

value of n'/n^0 may be assumed. Let $n'/n^0=\gamma$; by Eq. (5), $n'/n^0=1-\gamma$. From Eqs. (6) and (7),

$$Y_A{}^0=\gamma Y_A{}'+(1-\gamma)Y_A'$$

$$=\left[\frac{\gamma(\beta+p)}{P}+(1-\gamma)\right]\left[\frac{\gamma-(P-p)/\beta}{\gamma-1}\right]. \quad (8)$$

β is found by solution of this quadratic equation, all the other quantities being known; combination with (7) then gives Y_A'. This solves the problem, since the enrichment of A in the permeated gas is obtained from Y_A', and the area $\sum$ required to sustain an arbitrary rate of permeation is obtained from β.

Case II

Laminar flow (no mixing) occurs on either side of the film; the gas composition at any point on the low pressure side is given by the relative rates of permeation of the individual constituents at that point. From Eq. (1) there results

$$-dn_1=Q_A d\sum\left(P\frac{n_A}{n_A+n_B}-p\frac{dn_A}{dn_A+dn_B}\right) \quad (9)$$

and

$$-dn_B=Q_B d\sum\left(P\frac{n_B}{n_A+n_B}-p\frac{dn_B}{dn_A+dn_B}\right). \quad (10)$$

Let $x=dn_A/dn_B$, $i=n_A/n_B$, and $\alpha=Q_A/Q_B$; divide (9) by (10). Then

$$x=\alpha\left(P\frac{i}{i+1}-p\frac{x}{x+1}\right)\Big/\left(P\frac{1}{i+1}-p\frac{1}{x+1}\right). \quad (11)$$

Solution of this quadratic gives

$$x=f(i)=Ai+C+[A^2i^2+(2AC+\alpha)i+C^2]^{\frac{1}{2}}, \quad (12)$$

where

$$A=\tfrac{1}{2}[(1-\alpha)p/P+\alpha] \quad \text{and} \quad C=\tfrac{1}{2}[(1-\alpha)p/P-1].$$

Separation of variables in (12) can be achieved by noticing that $x=dn_A/dn_B=f(i)=f(n_1/n_B)$. Use of n_B and i as variables results in

$$\frac{dn_B}{n_B}=\frac{di}{f(i)-i}=\frac{di}{(A-1)i+C+[A^2i^2+2Bi+C^2]^{\frac{1}{2}}}, \quad (13)$$

where $2B=2AC+\alpha$. The right side of Eq. (13) can be rationalized by the introduction of a new variable t, defined by $i=\frac{1}{2}(C^2-t^2)/(At-B)$; the result is:

$$\ln\frac{n_B}{n_B{}^0}=-\int_{t^0}^{t}\frac{At^2-2Bt+AC^2}{(At-B)(t-\alpha+C)(t-C)}dt,$$

or

$$\ln\frac{n_B}{n_B{}^0}=R\ln\frac{t^0-B/A}{t-B/A}+S\ln\frac{t^0-\alpha+C}{t-\alpha+C}+T\ln\frac{t^0-C}{t-C}, \quad (14)$$

where $R=1/(2A-1)$, $S=[\alpha(A-1)+C]/(2A-1)[(\alpha/2)-C]$, and $T=1/(1-A-B/C)$; the superscript ^{0} designates entering gas. This solves the problem of determining the enrichment of the permeated gas for an arbitrary choice of n^0 $(=n_A{}^0+n_B{}^0)$, i^0, and i'; for a given $n_B{}^0=n^0/(1+i^0)$, n_B' is then found by Eq. (14), and n_A' is determined by $n_A'=i'n_B'$. The fraction of inlet gas which permeates is $(n^0-n')/n^0$; the fraction of A in the permeated gas is $(n_A{}^0-n_A')/(n^0-n')$.

It may be noted that the solution of Eq. (11) is particularly simple in the case $p=0$, for then $x=\alpha i$, or $dn_A/n_A=\alpha dn_B/n_B$, and

$$n_A/n_A{}^0=(n_B/n_B{}^0)^\alpha. \quad (15)$$

Computation of the permeation area required in Case II is somewhat involved. Assume there are no concentration gradients normal to the direction of gas flow. Use of Eqs. (12) and (13) in Eq. (11) then results in

$$\sum=-\frac{n_B{}^0}{Q_B}\int_{i^0}^{i'}\frac{\text{antiln}[n_B(i)/n_B{}^0]di}{[f(i)-i]\left[P\frac{1}{i+1}-p\frac{1}{f(i)+1}\right]}. \quad (16)$$

The integral on the right side is perhaps most conveniently evaluated numerically or graphically.

Because of the small magnitude of the permeability coefficients, it is probable that any practical design of a permeation process will involve flow conditions corresponding to Case II (laminar) rather than to Case I

TABLE II. Permeability of 0.0075-mm ethyl cellulose toward the constituents of air.

Gas	Permeability coefficient, Q, (moles cm^{-2} min.$^{-1}$ atmos.$^{-1}$) $\times 10^6$	Q_{gas}/Q_{N_2} Pure gas studies	Q_{gas}/Q_{N_2} Mass spectrometer
N_2	0.075	(1.0)	(1.0)
O_2	0.231	3.1	3.2
CO_2	1.07	14.3	—
A	0.170	2.3	3.1

TABLE II. Temperature dependence of permeability of 0.025-mm ethyl cellulose.

Gas	Temperature, °C	Permeability coefficient, Q, (moles cm^{-2} min.$^{-1}$ atmos.$^{-1}$) $\times 10^6$	Q_{gas}/Q_{N_2} at same temperature	Q_{gas}/Q_{O_2} at same temperature
N_2	30	0.025	(1)	0.31
	40	0.032	(1)	0.32
	50	0.045	(1)	0.35
O_2	30	0.081	3.3[a]	(1)
	40	0.100	3.2	(1)
	50	0.130	2.9	(1)
He	30	0.27	10.7	3.3
	40	0.33	10.2	3.3
	50	0.43	9.6	3.3
H_2	30	0.38	15.2	4.6
	40	0.46	14.4	4.6
	50	0.59	13.1	4.6

[a] A value of 3.4 has been obtained on other samples of the same film.

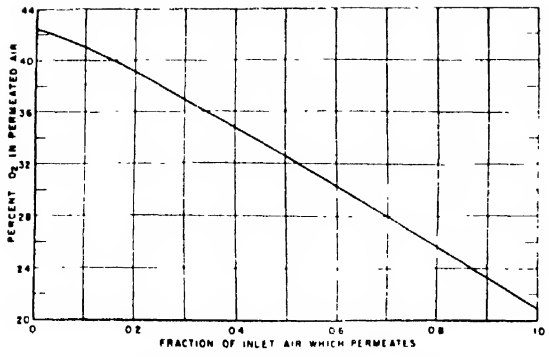


FIG. 2. Enrichment of air in a single stage of permeation. $P=8$, $p=1$, $Q_{O_2}/Q_{N_2}=3.4$.

(perfect mixing). It is interesting to note that, for a given set of conditions, Case II also results in a greater enrichment than does Case I. As an illustration, for the particular situation where normal air at 8 atmos. enters the stage and half of the entering air is permitted to permeate during the stage to the low pressure side at 1 atmos., the composition of the permeated air in Case I is 29.2 percent O_2, and in Case II 32.5 percent O_2. (A value of 3.4 was assumed for α.)

EXPERIMENTAL RESULTS

The permeability of various films toward O_2 and N_2, determined by experiments with the pure gases, is shown in Table I. The specific permeability, K, given is corrected for film thickness; $K=Qd$, where d is the thickness and Q is defined by Eq. (1). Where the comparisons are available, the specific permeability observed is seen to be in reasonable agreement with those previously reported. Of the materials investigated, natural rubber is outstanding in showing the highest permeability; its selectivity, however, as measured by the quantity $\alpha(\alpha=K_{O_2}/K_{N_2})$, is not as great as that of several other materials. Ethyl cellulose appears to satisfy best the joint requirements of high permeability and high selectivity.

Table II shows the permeability coefficients at 30°C of a laboratory-cast ethyl cellulose film toward the major constituents of air. Also shown is a comparison of the relative permeability (with respect to N_2) obtained from the pure gas studies and from the mass spectrometer analysis of air permeating through the film into a vacuum. The concentration of argon is increased along with that of O_2 in the permeated gas, though to a smaller extent than O_2. The permeability toward CO_2 is very high; in a multistage process to obtain high purity oxygen, it would probably be desirable to remove CO_2 after the first permeation stage.

The temperature dependence of the permeability of a 0.025-mm ethyl cellulose film toward several gases is illustrated in Table III. For each gas, the permeability increases with increasing temperature. The temperature dependence (and therefore the activation energy for permeation) is about the same for O_2, He, and H_2 but is greater for N_2. Van Amerongen[5] found a similar situation to hold for the same gases and natural rubber, N_2 having the highese activation energy for permeation and H_2, O_2, and He having lower and approximately equal activation energies. For the enrichment of air, a compromise has to be established between the use of higher temperatures to obtain higher permeability, and lower temperatures to obtain higher selectivity (α).

DISCUSSION

Primary consideration has been given above to a single stage of permeation. In the absence of a perfectly selective film, it is clear that to obtain gases of high purity it would be necessary to employ a multistage recycle system. The principles of such multistage arrangements have been discussed elsewhere.[7,8] Extension of the calculations given for a single stage to a multistage arrangement is straightforward but will not be considered here.

Since a practical permeation process probably would involve flow conditions leading to Eq. (14), calculation has been made of the enrichment of air to be expected from a single stage of permeation under a particular set of conditions. The high pressure has been chosen as 8 atmos., the low pressure as 1 atmos., and the enrichment factor $\alpha(=Q_{O_2}/Q_{N_2})$ as 3.4, the highest value which has been observed. Normal air (20.9 percent O_2) enters the stage. Under these conditions, Fig. 2 shows the composition of the permeated gas as a function of the fraction a of entering gas which permeates $[a=(n^0-n')/n^0]$. For negligible permeation ($a\doteq 0$), maximum enrichment of the permeated gas is observed, and the permeated gas contains 42.5 percent O_2. As $a\doteq 1$, the percent O_2 in the permeated gas approaches 20.9 percent, the inlet composition. For $a=0.5$, the permeated gas contains 32.5 percent O_2. Figure 2 is important for the computation of the net work required for the permeation, since the inlet air must be initially compressed by an imperfect compressor, and the energy of the nonpermeated, compressed gas leaving the stage must be utilized by some machine also of less-than-ideal efficiency.

It is clear from Table III that permeation processes may be of value for systems other than O_2-N_2. The usual separation of He from natural gas involves a relatively easy separation of He from CH_4 in one step, followed by a second step of more intense refrigeration to separate He from residual N_2. For the ethyl cellulose film of Table III, Q_{CH_4}/Q_{N_2} has been found to be about 1.5. The separation factor for He with respect both to N_2 and to CH_4 is quite large, which means that high enrichments can be achieved in each stage. Again, the permeability of ethyl cellulose to H_2 is much larger than to any other permanent gas with the exception of He. This has obvious applications, for example to the

[7] M. Benedict, Chem. Eng. Prog. 43, 41 (1947).
[8] K. Cohen, U. S. Naval Med. Bull. Supp. 1948, 6 (1948).

separation of H_2 from coke-oven gas or from the tail-gases of various hydrogenation processes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the suggestion of the problem and the continuing encouragement by Dr. H. H. Storch. Discussions with Dr. G. von Elbe and E. L. Clark have been most helpful. The authors wish also to thank Dr. R. A. Friedel and J. Sharkey for furnishing the mass-spectrometric analyses, and the companies listed in Table I for their generosity in furnishing samples of films.

D.C. Characteristics of Silicon and Germanium Point Contact Crystal Rectifiers. Part II. The Multicontact Theory*

V. A. Johnson, R. N. Smith,** and H. J. Yearian
Purdue University, Lafayette, Indiana

(Received August 29, 1949)

Part I has described the d.c. current-voltage characteristics obtained for typical Si and Ge crystal rectifiers, has given a summary of the published theories of the rectifier, and has shown that none of these theories accounts for the principal features of the observed characteristics. Part II describes the multicontact theory, an extension of the diode theory, and shows that this theory does account for the behavior of the observed characteristics. There also is described a graphical treatment, based upon the multicontact theory, which may be used in the analysis of experimental characteristics. The multicontact theory assumes that the contact potential at the metal-semiconductor surface varies from spot to spot. It is shown that an exponential distribution function for contact potentials may be used to account for the observed logarithmic slopes and that the observed temperature behavior of the characteristics may be explained by assuming that the distribution function varies with temperature in a manner dependent upon the nature of the contact.

1. INTRODUCTION

In Part I Yearian[1] has described the d.c. current-voltage characteristics obtained for typical silicon and germanium crystal rectifiers. A detailed knowledge of such characteristics and a complete theory of the rectifying behavior of such crystals have been the aim of extensive study in recent years because of the widespread use of silicon and germanium crystal rectifiers in microwave radar. Part I contains a summary of the heretofore published theories of the rectifier, and the discussion shows that none of these theories accounts for the salient features of the observed characteristics.

In Part II an extension of the diode theory, the multicontact theory, is described. It is shown that this theory does account for the essential properties of the observed characteristics. A graphical treatment based upon the multicontact theory is described and its application to the analysis of experimental characteristics is explained. The theory is used in a discussion of the observed change in characteristic for a given point contact with change in temperature of the contact region.

2. MULTICONTACT THEORY

It has been suggested (by Mott, Schottky, Lark-Horovitz, and Bethe) that the small slope of a typical observed log current *vs.* voltage characteristic may be due to a rapid variation in the properties of the blocking layer within the region of contact.[2] Since the thickness of the natural blocking layer is of the order of magnitude of the distance between impurity centers, it is quite probable that the contact potential at the surface of the semiconductor may vary from place to place. It is also likely that external impurities and adsorbed layers may produce fluctuations in contact potential.[3] The currents flowing across regions of low (negative or small positive) contact potential are essentially ohmic. In such regions the rate of change of current with applied voltage is much smaller than the exponential variation which occurs for higher contact potentials. The total current is the sum of the currents for regions of low and high contact potentials; hence, if the region of low contact potential is large enough in comparison to the region of high contact potential, the variation in total current with voltage rise is much less than the value predicted by pure diode theory for a contact with constant positive contact potential.

The multicontact theory assumes that the contact area consists of a number of spots, that the contact potential φ varies from spot to spot, and that the observed current is the sum of the contributions of

* This work was performed under OSRD Contract OEMsr-362 with Purdue University, and was reported in an NDRC Div. 14 Report dated August 14, 1943.

** Now at Boeing Aircraft Company, Seattle, Washington.

[1] H. J. Yearian, J. App. Phys. **21**, 214 (1950), referred to as Part I in the current discussion.

[2] Preliminary studies based upon the multicontact theory have been discussed by R. G. Sachs, NDRC Div. 14 Report No. 168, Purdue University (June 15, 1943); Phys. Rev. **69**, 682 (1946).

[3] For a different approach to the problem of how surface imperfections, foreign atoms on the surface, etc., may affect rectification, see John Bardeen, "Surface states and rectification at a metal semiconductor contact," Phys. Rev. **71**, 717 (1947).

the partial currents from these spots. The number of spots may be considered to be quite small with discretely separated contact potentials; in this case a graphical treatment of the summation is convenient. Or, a continuous distribution may be assumed such that the probability of finding a given contact potential between φ and $\varphi+d\varphi$ can be represented by a continuous probability density which is a function of φ, and so that the area of the spot having contact potential φ can also be represented as a probability function of φ. The spreading resistance in series with each spot is taken as $(2\sigma d)^{-1}$, where σ is the bulk conductivity of the semiconductor and d is the spot diameter. It follows that the spreading resistance is a function of φ determined by the dependence of spot area upon φ.

A graphical study shows that the addition of diode type currents for a group of spots corresponding to very few (three or four) different contact potentials produces a total current which is closely exponential, or linear on a semilogarithmic plot, over an appreciable voltage range. Any value of the logarithmic slope, less than e/kT, may be obtained by a suitable choice of the parameters governing the distribution. The graphical treatment is extended to provide a method whereby a total current curve may be rapidly synthesized to fit any given experimental characteristic.

One assumes that the contact consists of a number of small diodes, each with its own spreading resistance in series. These diodes are effectively in parallel, and so the total current is given by

$$i=\sum_j i_j, \tag{1}$$

where

$$i_j=j_0 a_j(\varphi_j)n_j(\varphi_j)\exp(-e\varphi_j/kT) \times[\exp\{(e/kT)(V_a-i_jR_{sj})\}-1] \tag{2}$$

is the current flowing through all spots for which the contact potential is φ_j. V_a is the applied voltage; j_0 denotes the current density due to the N free electrons (or holes) per unit volume in the semiconductor traveling with mean thermal velocity:

$$j_0=Ne\bar{v}/4=Ne(kT/2\pi m)^{\frac{1}{2}} \tag{3}$$

$a_j(\varphi_j)$ is the area of each spot having contact potential φ_j, and $n_j(\varphi_j)$ is the number of spots having contact potential φ_j. The spreading resistance R_{sj} for all spots with contact potential φ_j is calculated by considering the spreading resistances of such spots to be in parallel with one another. Then

$$R_{sj}=(\pi/a_j)^{\frac{1}{2}}(4\sigma n_j)^{-1}. \tag{4}$$

A template of a master curve (Fig. 1) to be used in the analysis of experimental characteristics, in the forward direction with $V_a>0.1$ volt, may be designed as follows: Using semilogarithmic paper with the same choice of scales as is made in plotting the experimental log$i$ *vs.* V_a characteristics, draw a straight line with logarithmic slope e/kT, i.e., the characteristic of a diode of single contact potential and zero spreading resistance. Choose an arbitrary value for the spreading resistance, compute the corresponding potential drops in the spreading resistance for a series of current values, and plot the characteristic of log$i$ *vs.* V_a, using the relation $V_a=V+R_{sj}i$. Mark on the template the point at which the iR_{sj} potential drop has a known value, such as 0.01 volt. This template may be used to reproduce the current-voltage curve for any diode of any φ_j; R_{sj} refers to the local series spreading resistance rather than to the equivalent spreading resistance of the whole contact.

In order to analyze a given experimental characteristic, by trial and error slide the template into several positions such that the sum of the partial currents so obtained closely approximates the given curve. Each partial current corresponds to a group of spots on the contact having the same contact potential, φ_j. The simplest numerical interpretation of the partial currents is made by assuming $a_j(\varphi_j)$ to be a constant with a reasonable value such as $a=10^{-10}$ cm^2. Then φ_j and $n_j(\varphi_j)$ may be evaluated as follows: Let i_j' represent the value of the partial current for which the potential drop in the spreading resistance is 0.01 volt, i.e., i_j' is the current at the index mark on the template. Then $R_{sj}=(100i_j')^{-1}$ ohm. Extend the low voltage straight

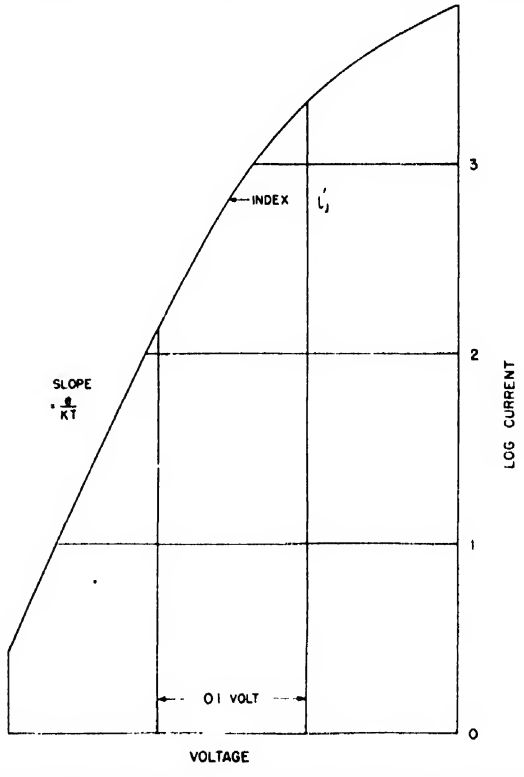


FIG. 1. Multicontact potential template such as is used in the analysis of experimental characteristics in the forward direction (applied voltage>0.1 volt). The index mark indicates the point on the characteristic for which the potential drop in the spreading resistance is 0.01 volt.

portion of the partial current back to find its intersection, i_{j0}, with the $V_a=0$ axis. Eq. (2) indicates that

$$i_{j0}=j_0 a n_j \exp(-e\varphi_j/kT). \qquad (2')$$

When σ and j_0 are known from resistivity and Hall effect measurements on the semiconductor, n_j may be computed by use of Eq. (4) and then Eq. (2') is used to determine φ_j. The product an_j represents the total area of the spots having contact potential φ_j. The choice of a different value for the area a will yield different values for the n_j, but the distribution of total areas (an_j) over the φ_j values will be much the same.

Figure 2 illustrates the application of the graphical method just described to the analysis of the characteristic obtained for contact in vacuum between a tungsten whisker and a germanium sample prepared in vacuum.[4] It is found that the observed characteristic may be treated as due to three groups of spots with $\varphi_j=0.1$, 0.4, and 0.8 volt, respectively.

Analogous templates and equivalent treatments may be designed for the analysis of experimental characteristics in the back direction and in the forward direction for $V_a<0.1$ volt.

It is to be noted that, in a construction of the type illustrated by Fig. 2, the chief contribution by a spot of contact potential φ_j to the approximately linear portion of the total current curve is made only over a range of V_a in the immediate neighborhood of φ_j. This affords an interpretation of any abnormal characteristic curve; e.g., if the curve of Fig. 2 were slightly concave upward in the region about $V_a=0.4$ volt, there would be a corresponding decrease in the total area corresponding to "B" spots. It is evident that, if the observed current-voltage curve is linear (on semilogarithmic plot) over a considerable range, the distribution of numbers of spots (for the simplest case of assumed constant area) must cover a similar range of φ and that, to obtain the linear total curve, the value of n_j must increase very rapidly with increasing φ_j.

The forward current intercept I_F is determined almost completely by the spots for which φ_j is greater than about 0.05 to 0.1 volt, and it may be seen that I_F is necessarily larger than the sum of the i_{j0} intercepts of the partial currents, $\sum_j i_{j0}$, where i_{j0} is given by Eq. (2). This sum is the contribution of these spots to the limiting diode current in the back direction. Thus one expects the observed diode back current I_B to be less than I_F. This is usually the case (Part I). The sum may be increased to give an I_B approaching I_F by introduction of spots of very low φ_j; this will affect I_F and the logarithmic slope α only slightly, but it will contribute noticeably to the observed diode current in the back direction. The ohmic component of the current in the back direction can be reproduced by any suitable distribution of spots having negative contact potentials; for such spots there is no barrier and only

[4] This figure and its analysis were supplied by Dr. R. M. Whaley of Purdue University; the authors thank him for this courtesy.

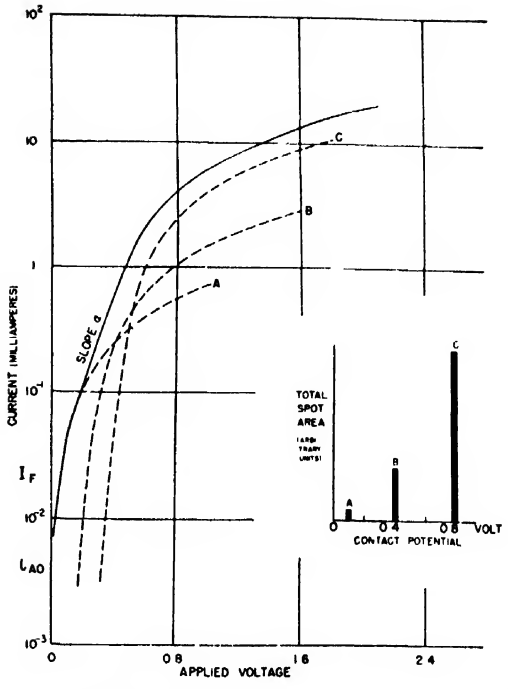


FIG. 2. Application of the multicontact theory to the analysis of an experimental characteristic. The solid line, indicating the experimental curve, is synthesized in a non-unique manner from three partial currents, A, B, and C, which correspond to three discrete contact potential groups. The insert indicates the distribution of spot area among the contact potentials 0.1, 0.4, and 0.8 volt, respectively.

R_{sj} need be considered. The observed back resistance usually can be accounted for by the existence of a very small total area of spots with negative φ_j. If the barrier lowering due to image force and tunneling effects (Part I) is considered, the spots of very small positive contact potenial will also be essentially ohmic in nature for current in the back direction.

3. EXPONENTIAL DISTRIBUTION

If a distribution function $P(\varphi)$ is introduced to describe the dependence of spot area upon contact potential, it is found that the theoretical current-voltage characteristic calculated with the use of such a distribution is essentially independent of the details of the function $P(\varphi)$. Only the tail of the distribution function at low contact potentials influences the shape of the characteristic in that part of the voltage range which is of practical interest. Hence it is not possible to determine the shape of the entire distribution function by use of an experimental characteristic. Since the theoretical characteristic depends only upon the tail of the distribution, one is free to choose a distribution function primarily on the basis of convenience in calculation. An exponential type distribution function is simple and convenient to use.

Hence it is assumed that

(1) The spot area depends upon φ_j according to

$$a(\varphi_j)=a_0 \exp(q\varphi_j). \qquad (5)$$

(2) The number of spots having contact potential φ_j is given by

$$n(\varphi_j)=n_0\exp(p\varphi_j). \tag{6}$$

p and q are parameters which may be chosen to fit experimental data.

(3) Insertions of Eqs. (5) and (6) into Eq. (4) yields for the spreading resistance due to all spots having contact potential φ_j:

$$R_{sj}=(4\sigma n_0)^{-1}(\pi a_0)^{-\frac{1}{2}}\exp(-p\varphi_j-q\varphi_j/2). \tag{7}$$

Equation (2) may now be written

$$i_j=n_0a_0j_0\exp\{(p+q-e/kT)\varphi_j\}\times[\exp(e/kT)(V_a-R_{sj}i_j)-1]. \tag{8}$$

If V_a is greater than about 0.1 volt, the -1 in the last bracket may be neglected, and Eq. (8) becomes

$$i_j=n_0a_0j_0\exp\{eV_a/kT-(e/kT-p-q)\varphi_j-R_{sj}i_j\}. \tag{9}$$

Figure 3 illustrates the construction of a total current curve by the summation of partial currents as given by Eq. (9) with $p=20$ volt^{-1} and $q=0$. Such a construction yields a total current curve which is straight over an appreciable range on a semilogarithmic plot and which is parallel to the envelope of the partial current curves. The logarithmic slope of the envelope can be shown to be

$$\alpha=\frac{p+q/2}{e/kT-q/2}\,\frac{e}{kT}\ \text{volt}^{-1};\quad (p+q)\gg e/kT. \tag{10}$$

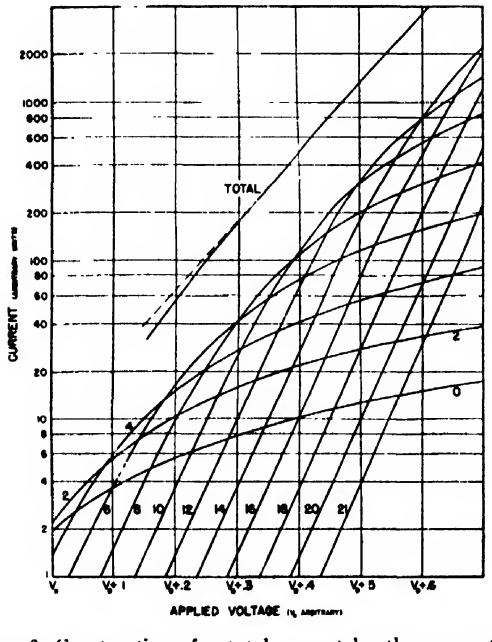


FIG. 3. Construction of a total current by the summation of partial currents. In this case the spot area is taken as constant and the number of spots having contact potential φ-volts is taken as $n_0\exp(20\varphi)$. The partial currents identified by $k=0, 2, \cdots$ are to be identified as passing through spots having $\varphi=\varphi_0+k/40$ volts, where φ_0 is an arbitrary contact potential.

By suitable choice of p and q, any value of α between 0 and e/kT may be obtained. The magnitude of i, but not the slope of the characteristic, depends upon constants of the material, such as σ, j_0, a_0, and n_0.

The assumptions used in the preceding analysis are the basis for the following theoretical treatment in which continuous distribution functions are employed. The dependence of spot area on contact potential is written

$$a(\varphi)=a_0\exp(q\varphi). \tag{11}$$

The number of spots having contact potential between φ and $\varphi+d\varphi$ is given by

$$n(\varphi)d\varphi=n_0\exp(p\varphi)d\varphi. \tag{12}$$

The spreading resistance due to all spots having contact potential φ is

$$R_s(\varphi)=(\pi/a_0)^{\frac{1}{2}}(4\sigma n_0)^{-1}\exp(-p\varphi-q\varphi/2). \tag{13}$$

Equations (11)–(13) may be used to set up an expression for the current in the forward direction. It is considered that the tendency is for diode type currents to flow through spots for which φ is positive and for ohmic currents to flow through spots for which φ is negative. These tendencies are restricted by the fact that the current density flowing through any group of spots is limited to the current density j_0 due to N electrons per unit volume traveling with thermal velocity. Reference to Eq. (2) for a diode type current shows that the current density cannot exceed j_0 if $\varphi>V$, but the current density will exceed j_0 if $\varphi<V$ unless some restriction is introduced. In order to conveniently express the total forward current, it is assumed that if V is large enough to produce the limiting current through a given group of spots, then further increase in V does not change the current. Since the current through a single spot is limited to $j_0a(\varphi)$, the corresponding potential drop in the spreading resistance in series with this spot is $j_0\{\pi a(\varphi)\}^{\frac{1}{2}}/4\pi\sigma$. This quantity has a small magnitude (usually less than 0.01 volt) and so the $iR_s(\varphi)$ drop may be neglected without serious error. V is assumed to be large enough ($\gg kT/e$) so that one may replace $\{\exp(eV/kT)-1\}$ by $\exp(eV/kT)$. Then the forward current ($V>0$) becomes

$$i(V)=j_0a_0n_0\exp(eV/kT)\times\int_V^\infty\exp\{-(e/kT-p-q)\varphi\}d\varphi+j_0a_0n_0\int_{-\infty}^V\exp\{(p+q)\varphi\}d\varphi \tag{14}$$

$$=\frac{j_0a_0n_0}{\alpha(e/kT-\alpha)}\,\frac{e}{kT}\exp(\alpha V),$$

where $\alpha=p+q$. By setting $V=0$ in Eq. (14) one ob-

TABLE I. Values obtained from study of the d.c. characteristics at various temperatures for germanium crystals in contact with tungsten whiskers. Columns 2, 3, 4, 5, and 7 give measured slopes, intercepts, and back resistances. Columns 6 and 7 permit checking the predicted relation $I_B = sI_F$. Columns 8, 9, and 10 list bulk properties of the germanium samples. Columns 11, 12, and 13 give calculated quantities characteristic of the exponential distribution used in applying the multicontact theory to these samples.

	1	2	3	4	5	6	7	8	9	10	11	12	13
Sample	T °K	α volt^{-1}	$s=\alpha kT/e$	I_F ma	I_B ma	sI_F ma	R_B ohms	σ mho/cm	Hall Constant cm^3/coulomb	j_0 10^4 amp./cm^2	n_0a_0 10^{-10} cm^2/volt	Minimum a_0 10^{-3} cm^2	n_0 volt^{-1}
$5I-d4$	243	22	0.46	0.0003	0.0004	0.00014	254000	28.2	64.5	4.43	0 813	0.35	0 023
	295	20	0.52	0.003	0.0032	0.0016	77000	23.5	61.5	5.09	5 74	1.34	0.043
	394	8.9–6.2	0.30–0.21	0.25–0.40	0.24	0.08	5800	16.8	58.5	6.18	291	72	0.040
	549	7.0–5.0	0.33–0.24	2.0 –3.0	0.3–0.5	0.29	90	18.7	23	17.8	607	15	0.41
$5I-d2$	286	7.4	0.18	0.20	0.025	0.04	1100	24.0	61.5	5.01	241	3 8	0.63
	436	6.3	0.24	1.0	0.35	0.24	2000	14.8	56.0	6.81	703	56	0.13
	505	4.9–2.8	0.21–0.12	2.0 –4.0	0.6	0.5	200	15.1	37	11.5	833	2.2	3.7
$9H-a$	298	7.8	0.20	0.80	0.013	0.16	32000	15.4	280	1.13	4420	400000	0.00011
	358	11.9	0.36	0.71	0.28	0.25	4980	12.3	277	1.25	4380	2600	0.017
	391	11.5	0.38	0.95	0.87	0.36	1670	11.2	275	1.32	5130	360	0.14
	426	9.0	0.33	2.5	3.0	0.8	495	10.6	217	1.68	8970	140	0.64
	468	6.6	0.27	5.5	7.1	1.5	145	11.0	105	2.20	12100	44	2.8
	521	7.0	0.31	10.2	7.8	3.2	19	15.4	37	4.17	11800	1.2	95
	574	4.4	0.22	22	?	4.8	7	27.0	16.5	9.41	8030	0.6	132

tains the forward current intercept

$$I_F = \frac{j_0 a_0 n_0}{\alpha(e/kT-\alpha)} \frac{e}{kT}. \tag{15}$$

The diode type current in the back direction is due to all spots having positive contact potentials. Current limitation need not be considered because of the smallness of the diode current for $V<0$, and for the same reason the $R_s i$ potential drop may be neglected. Then the diode current in the back direction ($V<0$) becomes

$$i(V) = j_0 a_0 n_0 \{\exp(eV/kT)-1\} \times \int_0^\infty \exp\{-(e/kT-p-q)\varphi\} d\varphi$$

$$= \frac{j_0 a_0 n_0}{e/kT-\alpha} \{\exp(eV/kT)-1\}. \tag{16}$$

The magnitude of the limiting current in the back direction is found by letting $V \to -\infty$; hence

$$I_B = j_0 a_0 n_0 (e/kT-\alpha)^{-1}. \tag{17}$$

The ohmic current in the back direction is due to all spots having negative contact potentials. The resistance in the back direction is given by

$$1/R_B = 4\sigma n_0 (a_0/\pi)^{\frac{1}{2}} \int_{-\infty}^0 \exp\{(p+q/2)\varphi\} d\varphi$$

$$= 4\sigma n_0 (a_0/\pi)^{\frac{1}{2}} (p+q/2)^{-1}. \tag{18}$$

The value of the quantity $n_0 a_0$ may be found directly from either Eq. (15) or Eq. (17); Eq. (15) is preferable because I_F is usually determined with greater accuracy than I_B. A consistency test for the theory is provided by the fact that the values of $n_0 a_0$ obtained by use of Eqs. (15) and (17) should agree within limits dependent upon the experimental errors in the values of I_F and I_B. These values of $n_0 a_0$ will be equal if

$$I_B = sI_F, \tag{19}$$

where s is the fractional slope $\alpha kT/e$. Table I shows corresponding values of I_B and sI_F, which are found to agree in order of magnitude. For the same reasons as given in the discussion on the discrete distribution of spots of various contact potentials, I_F is determined by only the spots for which the contact potential is greater than about 0.1 volt, whereas I_B is increased by the presence of spots of lower positive contact potential. Hence I_B may exceed the value given in Eq. (19). In addition, the effect of barrier lowering on the spots of low contact potential tends to increase the measured value of I_B.

The maximum value of $n_0 a_0^{\frac{1}{2}}$ may be calculated from Eq. (18) by setting $q=0$; the minimum value of $n_0 a_0^{\frac{1}{2}}$ is one-half of the maximum value and corresponds to setting $p=0$. A combination of $n_0 a_0$ and $n_0 a_0^{\frac{1}{2}}$ values yields values for a_0 and n_0; such values are listed in Table I.

In general the values of $n_0 a_0$ are reasonable, i.e., they are small compared to the total whisker area of 10^{-6} to 10^{-5} cm^2. There is a noticeable tendency for $n_0 a_0$ values to increase with rising temperature. At the lower temperatures the a_0 values tend to exceed the $n_0 a_0$ values, thus yielding fractional values for n in the neighborhood of zero contact potentials. Such a result is contradictory to the basic assumption of a statistical distribution of a very large number of spots.

One possible explanation of the small n_0 values is that the assumed exponential distribution may not be valid in the region of negative contact potential. The back resistance determines the value of

$$\int_{-\infty}^0 n(\varphi)\{a(\varphi)\}^{\frac{1}{2}} d\varphi$$

and, for any assumed distribution, the value of $n_0 a_0^{\frac{1}{2}}$.

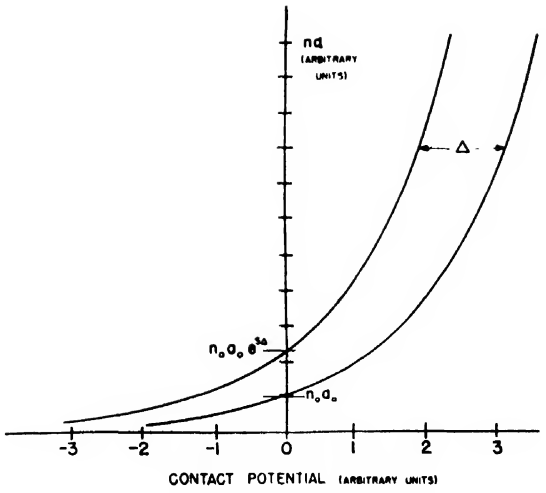


FIG. 4. Effect of temperature change upon the distribution function for "welded" contacts. Ordinates represent the contact areas corresponding to various contact potentials.

A distribution falling off more rapidly with increasing negative φ than an exponential distribution leads to a larger value of $n_0a_0^3$ for a fixed n_0a_0. This produces a smaller calculated value of a_0 and hence a larger value of n_0, as is desired.

A second explanation is that the effective conductivity of the small ohmic spots is less than the measured bulk conductivity of the sample. Then a_0 is less in proportion as σ^2 drops below the bulk value. Weisskopf[5] has advanced the hypothesis that the small spots with negative contact potential may be "overloaded" and that "overloading" decreases the effective conductivity. The exponential dependence of spot area on contact potential may lead, for negative φ, to spot diameters equal to or smaller than the electronic mean-free-path in the semiconductor; Lark-Horovitz has pointed out that in such a case a reduction in conductivity is to be expected.

A section of Part I described the effect of image force and tunneling in producing a barrier lowering in the point contact rectifier. A quantitative introduction of barrier lowering into the diode rectifier theory has been made by Courant.[6] Since the multicontact theory is based upon the diode theory, calculations have been made to ascertain the effects of introducing barrier lowering into the multicontact theory. For the forward direction the changes are small and may be summarized as follows: (1) at high voltage the logarithmic slope is constant at the same value given by the multicontact theory without barrier lowering, but at lower voltages (less than about 0.4 volt) the logarithmic slope with lowering considered decreases slightly with decreasing voltage; (2) for a fixed n_0a_0 value the current calculated with barrier lowering considered is greater than the current at the same voltage without lowering; the factor varies from 1.0 to 1.2, increasing with increasing α; (3) for a fixed n_0a_0 value one obtains a higher I_F intercept on the theoretical curve including barrier lowering; I_F is higher by about 5 percent to 20 percent, depending upon α. Another way of stating this result is to say that, for an I_F value taken from a measured characteristic, one calculates on the basis of barrier lowering a value for n_0a_0 that is from 5 percent to 20 percent lower than he would obtain by neglecting barrier lowering.

Since the effect of tunneling alone is to give rectification in the opposite direction from the observed direction, one would expect the inclusion of tunneling, through barrier lowering, in the multicontact theory to give rather marked effects in the back direction characteristic. Such is found to be the case; both the ohmic and diode contributions as calculated without barrier lowering are altered. The back resistance corresponding to the ohmic current is considerably reduced and hence the ohmic current considerably increased. The diode contribution is markedly changed; the extent of the change depends upon the electron concentration with the change greatest for the highest concentration. The tunneling produces an exponentially increasing current that, at higher voltages, has a logarithmic slope equal to or even greater than the forward slope. This exponential current tends to cover up the saturation current, denoted by I_B. The tendency is for the ohmic current to dominate at the lower back voltages; the extent of barrier lowering determines the voltage at which the exponential current surpasses the ohmic current.

The barrier lowering calculations of Courant assume an electron concentration of 5×10^{18} cm^{-3}; the use of these results in the multicontact theory yields calculated back currents too high to agree with experiment. The agreement would be improved by noting that the actual electron concentrations in germanium and silicon point contact rectifiers are one or two orders of magnitude less than that assumed by Courant, and hence the effect of barrier lowering would not be as great as calculated by Courant.

4. TEMPERATURE BEHAVIOR

In Part I a description was given of the changes in characteristic observed for a given contact as the temperature is varied. Sets of characteristics for different temperatures were obtained both for "free" contacts between tungsten whiskers and germanium and for "welded" contacts between whisker and germanium in General Electric cartridges. The temperature behavior of α is dependent upon the type of contact. For free contacts, α decreases with increasing temperature in such a way that the fractional slope is approximately constant. For welded contacts, the value of α is prac-

[5] V. F. Weisskopf, Radiation Laboratory Report 133 (May, 1943).

[6] E. Courant, NDRC Div. 14, Cornell University Report, Contract OEMsr-429 (May 17, 1943); Phys. Rev. 69, 684 (1946). A report of Courant's results is given in a paper by W. E. Meyerhof, Phys. Rev. 71, 727 (1947).

tically independent of temperature. Figures of Part I show that the back resistance R_B decreases with increasing temperature and that the forward intercept I_F and limiting diode current in the back direction, I_B, increase with rising temperature. The plot of $\log R_B$ against temperature can be represented by a straight line of negative slope; the plots of $\log I_F$ and $\log I_B$ against T can be approximated by straight lines of positive slope.

Equation (14) shows that the forward logarithmic slope α is equal to the sum of the distribution parameters p and q, and so α determines the shape of the distribution. The constancy of α with changing temperature for the "welded" contacts suggests that the shape of the distribution is independent of T for these contacts, but the dependence of R_B, I_F, and I_B on T implies that the distribution has some type of temperature dependence. Assume that, although the shape of the distribution curve is unchanged, the whole curve is shifted by an amount Δ depending upon the temperature change. Let φ' be contact potential, at the new temperature, corresponding to the ordinate of the distribution curve at φ at the original temperature (see Fig. 4). Then

$$\varphi' = \varphi - \Delta. \tag{20}$$

Introducing this assumption into Eqs. (14), (16), and (18) yields

$$I_F' = \frac{j_0' a_0 n_0}{\alpha(e/kT' - \alpha)} \frac{e}{kT'} \exp(\alpha\Delta) \approx I_F \exp(\alpha\Delta) \tag{21}$$

$$I_B' = j_0' a_0 n_0 (e/kT' - \alpha)^{-1} \exp(\alpha\Delta) \approx I_B \exp(\alpha\Delta) \tag{22}$$

$$R_B' = (\pi/a_0)^{\frac{1}{2}} (4\sigma' n_0)^{-1} (\alpha - q/2) \exp\{-(\alpha - q/2)\Delta\}$$
$$\approx R_B \exp\{-(\alpha - q/2)\Delta\}. \tag{23}$$

The introduction of the shift Δ makes it possible to treat n_0 and a_0 as temperature independent in Eqs. (21), (22), and (23). If j_0 is taken as proportional to $T^{\frac{1}{2}}$, the temperature dependence predicted by Eq. (21) is

$$(T')^{\frac{1}{2}}(e/kT' - \alpha) I_F' \sim \exp(\alpha\Delta).$$

Plots against T of $\log I_F'$ and $\log\{(T')^{\frac{1}{2}}(e/kT' - \alpha) I_F'\}$ give lines of equal slopes and so justify the use of the simpler expression. As shown in Part I, the plots of $\log I_F'$, $\log I_B'$, and $\log R_B'$ against T are quite straight over a considerable temperature range. This suggests the relation

$$\Delta = B(T' - T) \tag{24}$$

that is, the distribution curve is shifted laterally by an amount proportional to the temperature change. A change in temperature shifts the distribution so that the total area corresponding to negative contact potential increases with rising temperature. The value of B may be found from the known values of α and of the slopes of $\log I_F'$ *vs.* T and $\log I_B'$ *vs.* T lines. For the welded samples, $B = 1.1$–1.2 millivolts per degree as determined from the I_F plot and $B = 1.8$–2.0 mv per degree from the I_B plot.

The foregoing discussion does not apply to free contacts. Since the value of α decreases with rising temperature, the shape of the distribution curve changes and p and q are functions of temperature. The approximately constant value of the fractional slope implies that α, p, and q are inversely proportional to T. In order to explain the temperature behavior of the characteristics for free contacts, one must assume that the distribution curve for φ is not only shifted but is also flattened as the temperature rises.

5. SUMMARY

In conclusion, one can say that (1) the multicontact theory accounts for the frequently observed forward logarithmic slopes much less than e/kT, and for the diode and ohmic components of current in the back direction, (2) a distribution of spots with discrete contact potentials can readily be found such that it will yield any given observed characteristic; although such a discrete distribution is non-unique, it is useful in interpretation of the given characteristic, (3) an exponential distribution function, with parameters determined by experimental values, is satisfactory and convenient, and (4) the observed temperature behavior may be accounted for by assuming that the distribution function varies with temperature in a manner dependent on the nature of the contact. It should be noted that the treatment just described is incomplete in the sense that the constants of the distribution may be found empirically from the observed characteristics, but the constants cannot be determined by independent physical reasoning.

Analysis of a Sampling Servo Mechanism

KENNETH S. MILLER* AND RALPH J. SCHWARZ†
Columbia University, New York, New York
(Received October 3, 1949)

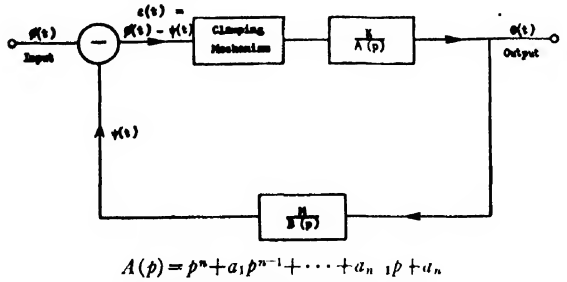
An analysis of a sampling servo mechanism with an error-clamping device and linear forward and return paths is given. The method leads to a determination of the continuous output of the system in terms of quadratures and gives explicitly the value of the output at the discrete sampling instants. Frequently this is all that is needed in order to study the output variation with time. The question of stability is discussed and a criterion for testing the stability of the system is given in terms of the system parameters.

I. INTRODUCTION

IN a large class of servo mechanisms, error data are supplied continuously, the error being a measure of the instantaneous deviation of some function of the output from the input. In another important class of servos, called sampling servo mechanisms,[1,2] error data are supplied at discrete, equally spaced instants only. The servo mechanism is then given no information about the error in the intermediate periods. Such systems may be analyzed by using the theory of filters with pulsed data.[3]

In many applications, the error data, supplied only at discrete sampling instants, are passed through a device whose output at any time is equal to the input at the preceding sampling instant. Such a mechanism is called a "clamping device." The output of the error-clamping device is thus constant for a period equal to the interval between sampling instants. This interval will be called the sampling period, T_s. Sampling occurs at the instants $t=0, T_s, 2T_s, \cdots, NT_s, \cdots$.

The mathematical analysis below is carried out by employing the "one-sided Green's function." Using this and associated concepts, explicit relations are derived for the error, response to a step input, and asymptotic value of the output; further, a stability criterion is formulated. The output is a continuous function of time, and an integral representation is obtained. An explicit formula, however, can be given for the output at the sampling instants.



$A(p)=p^n+a_1p^{n-1}+\cdots+a_{n-1}p+a_n$
$B(p)=p^m+b_1p^{m-1}+\cdots+b_{m-1}p+b_m$

FIG. 1. Sampling servo system.

II. MATHEMATICAL ANALYSIS

1. Introduction

The feed-back system which is to be operated as a sampling servo mechanism in the sense of the definition given in Section I is illustrated in Fig. 1. $A(p)$ and $B(p)$ are polynomials in p with constant coefficients. p is the differential operator d/dt. Explicitly,

$$A(p)=p^n+a_1p^{n-1}+\cdots+a_{n-1}p+a_n \tag{1}$$

$$B(p)=p^m+b_1p^{m-1}+\cdots+b_{m-1}p+b_m. \tag{2}$$

K and M are constants and

$$R=KM$$

is defined as the "loop gain."

It will be assumed that the system is initially unexcited. This implies that

$$\psi^{(\alpha)}(0)=0, \quad \alpha=0, 1, \cdots, m+n-1 \tag{3}$$

where $\psi(t)$ is defined in Fig. 1.

In accordance with Fig. 1,

$$A(p)\theta(t)=K\cdot\epsilon(t) \tag{4}$$

and

$$B(p)\psi(t)=M\cdot\theta(t).$$

Hence

$$A(p)B(p)\psi(t)=MK\cdot\epsilon(t)\equiv R\cdot\epsilon(t). \tag{5}$$

Since the initial conditions (3) have been specified, the inverse operator $[A(p)B(p)]^{-1}$ is well defined. In fact,

$$\psi(t)=\int_0^t G(t,\zeta)[R\cdot\epsilon(\zeta)]d\zeta \tag{6}$$

where $G(t,\zeta)$ is the "one-sided Green's function" for the incompatible system[4]

$$\left.\begin{aligned} &A(p)B(p)\psi(t)=0 \\ &\psi^{(\alpha)}(0)=0, \quad \alpha=0, 1, \cdots, m+n-1 \end{aligned}\right\} \tag{7}$$

---

* Present address: School of Mathematics, Institute for Advanced Study, Princeton, New Jersey.

† Department of Electrical Engineering.

[1] L. A. MacColl, *Fundamental Theory of Servomechanisms* (D. Van Nostrand Company, Inc., New York, 1945), Chapter 10.

[2] R. C. Oldenbourg and H. Sartorius, *The Dynamics of Automatic Controls*, translated and edited by H. L. Mason (American Society of Mechanical Engineers, New York, 1948), Chapter 5.

[3] James, Nichols, and Philips, *Theory of Servomechanisms*, Vol. 25 of M.I.T. Radiation Laboratory Series (McGraw-Hill Book Company, Inc., New York, 1947), Chapter 5.

[4] E. L. Ince, *Ordinary Differential Equations* (Dover Publications, New York, 1944), p. 254 ff.

and is given by

$$G(t,\zeta)=(-1)^{m+n-1}\frac{\begin{vmatrix}\phi_1(t) & \phi_2(t) & \cdots & \phi_{m+n}(t)\\ \phi_1(\zeta) & \phi_2(\zeta) & \cdots & \phi_{m+n}(\zeta)\\ \phi_1'(\zeta) & \phi_2'(\zeta) & \cdots & \phi_{m+n}'(\zeta)\\ \cdots & \cdots & \cdots & \cdots\\ \phi_1^{(m+n-2)}(\zeta) & \phi_2^{(m+n-2)}(\zeta) & \cdots & \phi_{m+n}^{(m+n-2)}(\zeta)\end{vmatrix}}{\begin{vmatrix}\phi_1(\zeta) & \phi_2(\zeta) & \cdots & \phi_{m+n}(\zeta)\\ \cdots & \cdots & \cdots & \cdots\\ \phi_1^{(m+n-1)}(\zeta) & \phi_2^{(m+n-1)}(\zeta) & \cdots & \phi_{m+n}^{(m+n-1)}(\zeta)\end{vmatrix}}. \tag{8}$$

In this determinant, $\phi_\alpha(t)$, $\alpha=1, 2, \cdots, m+n$, are $m+n$ linearly independent solutions of $A(p)B(p)\psi(t)=0$.

It should be noted that no loss of generality results by considering R a constant. For if R were a polynomial in p,

$$R\equiv R(p)=p^q+c_1p^{q-1}+\cdots+c_{q-1}p+c_q \tag{9}$$

then

$$R(p)\epsilon(t)\equiv c_q\epsilon_{N-1}$$

in any sampling interval $(N-1)T_s<t\leqq NT_s$ since $\epsilon(t)$ is a "staircase function" (see Section I). Also, it is permissible to use a "one-sided Green's function" instead of the more familiar "two-sided" one[4] since symmetric properties of the integral operator of Eq. (6) are of no interest in this paper.

Since the operator $A(p)B(p)$ has constant coefficients, the linearly independent solutions $\phi_\alpha(t)$ are of the form

$$\phi_\alpha(t)=t^re^{st}, \quad \alpha=1, 2, \cdots, m+n. \tag{10}$$

It will now be assumed that the $m+n$ roots, v_α, $\alpha=1, 2, \cdots, m+n$ of the characteristic equation for $A(p)B(p)$ are all distinct and non-zero. The analysis of cases not in this category will be readily apparent from the treatment given here.

With this assumption,

$$\phi_\alpha(t)=e^{v_\alpha t}, \quad \alpha=1, 2, \cdots, m+n. \tag{11}$$

Substituting these values of $\phi_\alpha(t)$ in the Green's function, Eq. (8), yields

$$G(t,\zeta)=(-1)^{m+n}\sum_{\beta=1}^{m+n}(-1)^\beta e^{v_\beta t}e^{-v_\beta\zeta}W_\beta W^{-1} \tag{12}$$

where W is the Vandermonde determinant[5]

$$W=\begin{vmatrix}1 & 1 & \cdots & 1\\ v_1 & v_2 & \cdots & v_{m+n}\\ \cdots & \cdots & \cdots & \cdots\\ v_1^{m+n-1} & v_2^{m+n-1} & \cdots & v_{m+n}^{m+n-1}\end{vmatrix}=\prod_{\substack{\alpha,\beta=1\\ \alpha>\beta}}^{m+n}(v_\alpha-v_\beta) \tag{13}$$

and W_β is the minor of the element in the last row and β^{TH} column of W.

$$W_\beta=\prod_{\substack{i,j=1\\ i>j\\ i,j\neq\beta}}^{m+n}(v_i-v_j), \quad \beta=1, 2, \cdots, m+n. \tag{14}$$

2. The Linear Recurrent Relation for $\epsilon(t)$

Since $\epsilon(t)$ is a staircase function, the value of $\psi(t)$ at the sampling instant $t=NT_s$ is

$$\psi(NT_s)\equiv\psi_N=R\sum_{\alpha=0}^{N-1}\epsilon_\alpha\int_{\alpha T_s}^{(\alpha+1)T_s}G(NT_s,\zeta)d\zeta \tag{15}$$

where $\epsilon(t)=\epsilon_\alpha$ for $\alpha T_s<t\leqq(\alpha+1)T_s$. Introducing the notation

$$\left.\begin{aligned}&\omega_\beta=e^{v_\beta T_s}\\ &X_\beta=(-1)^{m+n+\beta}W_\beta W^{-1}(1-\omega_\beta^{-1})v_\beta^{-1}\end{aligned}\right\} \quad \beta=1, 2, \cdots, m+n, \tag{16}$$

Eq. (15) becomes, after a few simple algebraic reductions,

$$\psi_N=R\sum_{\alpha=0}^{N-1}\epsilon_\alpha\sum_{\beta=1}^{m+n}X_\beta\omega_\beta^{N-\alpha}, \quad \beta=1, 2, \cdots, m+n. \tag{17}$$

Letting

$$Z_{j,\nu}=\sum_{\alpha=0}^{\nu-1}\epsilon_\alpha\omega_j^{-\alpha}, \quad j=1, 2, \cdots, m+n, \tag{18}$$

yields

$$Z_{j,N+q}=Z_{j,\nu}+\sum_{\beta=0}^{q-1}\epsilon_{N+\beta}\omega_j^{(N+\beta)}, \quad \begin{matrix}j=1, 2, \cdots, m+n\\ q=1, 2, \cdots.\end{matrix} \tag{19}$$

The second term of Eq. (19) does not involve any summation "up to N." It is desired to obtain a relation among the ψ_N's which is free of $Z_{j,N}$. Toward this end the ψ_{N+q}'s are considered.

$$\psi_{N+q}=R\sum_{\alpha=1}^{m+n}X_\alpha\omega_\alpha^{N+q}Z_{\alpha,N}+R\sum_{\alpha=1}^{m+n}X_\alpha\sum_{\beta=0}^{q-1}\epsilon_{N+\beta}\omega_\alpha^{q-\beta}. \tag{20}$$

Introducing the $m+n+1$ multipliers λ_q and summing

[5] G. Kowalewski, *Einfuhrung in die Determinantentheorie* (Chelsea Publishing Company, New York, 1948), p. 38.

the products $\lambda_q\psi_{N+q}$ up to $m+n$ yields

$$\sum_{q=0}^{m+n}\lambda_q\psi_{N+q}=R\sum_{q=0}^{m+n}\lambda_q\sum_{\alpha=1}^{m+n}X_\alpha\omega_\alpha{}^{N+q}Z_{\alpha,N}$$

$$+R\sum_{q=0}^{m+n}\lambda_q\sum_{\alpha=1}^{m+n}X_\alpha\sum_{\beta=0}^{q-1}\epsilon_{N+\beta}\omega_\alpha{}^{q-\beta}. \quad (21)$$

The λ_α, $\alpha=0, 1, \cdots, m+n$ will be so determined that the first member of Eq. (21) on the right vanishes without all the λ_α being identically zero. It is seen that

$$\sum_{q=0}^{m+n}\lambda_q\omega_\alpha{}^q=0, \quad \alpha=1, 2, \cdots, m+n \quad (22)$$

is a sufficient condition. If $\lambda_{m+n}=1$, then Eq. (22) becomes a system [Eq. (23)] of $(m+n)$ equations for the $(m+n)$ unknowns $\lambda_0, \lambda_1, \cdots, \lambda_{m+n-1}$:

$$\sum_{q=0}^{m+n-1}\lambda_q\omega_\alpha{}^q=-\omega_\alpha{}^{m+n}, \quad \alpha=1, 2, \cdots, m+n. \quad (23)$$

Since the determinant of this system is again a Vandermonde determinant, it is $\neq 0$ since $\omega_\alpha\neq\omega_\beta$ (and $\alpha\neq\beta$) by hypothesis. Hence the system of equations is consistent. Since $\lambda_{m+n}=1\neq 0$, not all the λ_α vanish. With this choice of the λ's, Eq. (21) reduces (after some algebraic manipulations) to

$$\epsilon_{N+m+n}+\sum_{\beta=0}^{m+n-1}\{\lambda_\beta+R\sum_{q=\beta+1}^{m+n}\sum_{\alpha=1}^{m+n}\lambda_qX_\alpha\omega_\alpha{}^{q-\beta}\}\epsilon_{N+\beta}$$

$$=\sum_{q=0}^{m+n}\lambda_q\phi_{N+q}. \quad (24)$$

Equation (24) is an equation for the ϵ_N's with constant coefficients. It is not a difference equation in the usual sense since it is valid only at the sampling instants $0, T_s, 2T_s, 3T_s, \cdots$. Hence, strictly speaking, Eq. (24) is a linear recurrence formula.[6]

From Eq. (23) it is seen that the λ's are the indicated coefficients of the algebraic equation

$$x^{m+n}+\lambda_{m+n-1}x^{m+n-1}+\cdots+\lambda_1x+\lambda_0=0, \quad \lambda_{m+n}=1 \quad (25)$$

whose roots are ω_α, $\alpha=1, 2, \cdots, m+n$.

If Eq. (24) is written in the compact form

$$\sum_{\alpha=0}^{m+n}A_\alpha\epsilon_{N+\alpha}=\sum_{q=0}^{m+n}\lambda_q\phi_{N+q} \quad (26)$$

where

$$A_{m+n}=1, \quad A_\beta=\lambda_\beta+R\sum_{q=\beta+1}^{m+n}\sum_{\alpha=1}^{m+n}\lambda_qX_\alpha\omega_\alpha{}^{q-\beta}, \quad (27)$$

$$\beta=0, 1, \cdots, m+n-1,$$

it is seen that the solution of the homogeneous equation

$$\sum_{\alpha=0}^{m+n}A_\alpha\epsilon_{N+\alpha}=0 \quad (28)$$

is

$$\epsilon_N=\sum_{\beta=1}^{m+n}P_\beta r_\beta{}^N \quad (29)$$

where the r_β are the roots of the algebraic equation

$$\sum_{\alpha=0}^{m+n}A_\alpha r^\alpha=0 \quad (30)$$

and P_β are periodic constants of period T_s. Since Eq. (24) is a linear recurrence formula and not a difference equation, the P_β are precisely constants in the ordinary sense. In writing Eq. (29), it has been tacitly assumed that all the roots of Eq. (30) are distinct. If this is not the case, the obvious modification

$$\epsilon_N=\sum_{\beta=1}^{k}[P_\beta N^{\beta-1}]r_1{}^N+\sum_{\alpha=k+1}^{m+n}P_\alpha r_\alpha{}^N$$

is made, where r_1 is a root of multiplicity k.

The particular integral $h(N)$ corresponding to the non-homogeneous equation can be obtained by the usual method of undetermined coefficients.[7] Hence the solution of Eq. (26) is

$$\epsilon_N=\sum_{\beta=1}^{m+n}P_\beta r_\beta{}^N+h(N). \quad (31)$$

The constants P_β in the solution [Eq. (31)] corresponding to the boundary conditions of Eq. (3) can readily be obtained in the following manner.

At the time $t=0$, $\psi(0)=0$; hence

$$\phi(0)=\epsilon(0) \quad \text{or} \quad \phi_0=\epsilon_0. \quad (32)$$

From Eq. (15)

$$\psi_1=R\epsilon_0\int_0^{T_s}G(T_s, \zeta)d\zeta. \quad (33)$$

Since ϵ_0, $\int_0^{T_s}G(T_s, \zeta)d\zeta$ and R are known, ψ_1 is known. ϕ_1 is known since the input $\phi(t)$ is a given function. Since at the sampling instants

$$\epsilon_N=\phi_N-\psi_N, \quad (34)$$

in particular

$$\epsilon_1=\phi_1-\psi_1 \quad (35)$$

and hence ϵ_1 is known. Proceeding in this pedestrian fashion, $\epsilon_0, \epsilon_1, \cdots, \epsilon_{m+n-1}$ are calculated. From Eq. (29)

$$\epsilon_\alpha=\sum_{\beta=1}^{m+n}P_\beta r_\beta{}^\alpha+h(\alpha), \quad \alpha=0, 1, \cdots, m+n-1. \quad (36)$$

The determinant of the P_β is non-zero. (It is a Vander-

[6] T. Fort, *Finite Differences* (Oxford University Press, London, 1948), p. 115.

[7] T. Fort (reference 6, p. 126).

monde determinant if the r_β are distinct and non-zero.) Hence the system (36) may be solved for P_β, $\beta=1, 2, \cdots, m+n$.

3. Stability of the Sampling Servo Mechanism

The servo system illustrated in Fig. 1 will be said to be stable if $\epsilon_N \to 0$ as $N \to \infty$ for any bounded input $\phi(t)$ of finite duration. If $\phi^*(t)$ is such a function, $\phi_N{}^*=0$ and $h(N)=0$ for N sufficiently large. Therefore, Eq. (31) may be written as

$$\epsilon_N = \sum_{\beta=1}^{m+n} P_\beta r_\beta{}^N. \tag{37}$$

It is evident from Eq. (37) that a necessary and sufficient condition that $\epsilon_N \to 0$ as $N \to \infty$ is that $|r_\beta|<1$. In other words, all the roots r_β, $\beta=1, 2, \cdots, m+n$ must lie inside the unit circle in the complex domain. Since the unit circle in the z plane can be transformed into the left half-plane of the complex w plane, the Nyquist criterion can be applied to determine the stability of the system without actually calculating the roots.

4. Determination of the Output $\theta(t)$

To determine the output $\theta(t)$, it is noted from Fig. 1 that

$$A(p)\theta(t)=K\epsilon(t). \tag{38}$$

Also, from the feed-back loop

$$B(p)\psi(t)=M\theta(t). \tag{39}$$

Since $\psi^{(\alpha)}(0)=0$, $\alpha=0, 1, \cdots, m+n-1$ by Eq. (3), it follows that

$$\theta^{(\alpha)}(0)=0, \quad \alpha=0, 1, \cdots, n-1. \tag{40}$$

Hence the "Green's function method" is applicable and

$$\theta(t)=K\int_0^t g(t, \zeta)\epsilon(\zeta)d\zeta \tag{41}$$

where $g(t, \zeta)$ is a one-sided Green's function.

If it is desired to compute $\theta(t)$ at the sampling instants nT_s, a method exactly analogous to the one discussed for $\epsilon(t)$ leads to the equation

$$\sum_{q=0}^{n} l_q\theta_{N+q}=M\sum_{q=0}^{n} l_q\sum_{\alpha=1}^{n} Y_\alpha\sigma_\alpha{}^q\sum_{\beta=0}^{q-1}\epsilon_{N+\beta}\sigma_\alpha{}^{-\beta} \tag{42}$$

where

$$\sigma_\alpha=e^{u_\alpha T_s}, \tag{43}$$

$u_1, u_2, \cdots, u_n$ are the distinct non-zero roots of the algebraic equation

$$x^n+a_1x^{n-1}+\cdots+a_{n-1}x+a_n=0, \tag{44}$$

$$Y_\alpha=(-1)^{n+\alpha}\Delta_\alpha\Delta^{-1}(1-\sigma_\alpha{}^{-1})u_\alpha{}^{-1}, \quad \alpha=1, 2, \cdots, n, \tag{45}$$

$$\Delta=\prod_{\substack{i,j=1\\ i>j}}^{n}(u_i-u_j), \tag{46}$$

$$\Delta_\alpha=\prod_{\substack{i,j=1\\ i>j\\ i,j\neq\alpha}}^{n}(u_i-u_j), \quad \alpha=1, 2, \cdots, n, \tag{47}$$

and l_α are the indicated coefficients of the algebraic equation

$$x^n+l_{n-1}x^{n-1}+\cdots+l_1x+l_0=0, \quad l_n\equiv 1, \tag{48}$$

whose roots are $\sigma_1, \sigma_2, \cdots, \sigma_n$. Hence if $f(N)$ is any particular solution of the non-homogeneous Eq. (42),

$$\theta_N=\sum_{\alpha=1}^{n} Q_\alpha\sigma_\alpha{}^N+f(N). \tag{49}$$

The constants Q_α, $\alpha=1, 2, \cdots, n$ can be obtained in the same manner in which the P_β were determined in Section II.2.

5. Response to a Step Input

If the input $\phi(t)$ to a stable system is given by $u(t)$ where $u(t)$ is the unit step function defined by

$$\left.\begin{aligned} u(t-\alpha)&=0 \quad t<\alpha\\ u(t-\alpha)&=1 \quad t>\alpha \end{aligned}\right\}, \tag{50}$$

then

$$\phi_0=\phi_1=\cdots=\phi_N=\cdots=1 \tag{51}$$

and

$$\sum_{q=0}^{m+n}\lambda_q\phi_N=\sum_{q=0}^{m+n}\lambda_q. \tag{52}$$

Hence,

$$h(N)=[\sum_{q=0}^{m+n}\lambda_q][\sum_{\alpha=0}^{m+n}A_\alpha]^{-1}\equiv H \tag{53}$$

is a particular solution of Eq. (26).

As $N\to\infty$, $\epsilon_N\to H$ since $|r_\beta|<1$, $\beta=1, 2, \cdots, m+n$. Hence it is seen that the system does not close out the error arising from a step function input. In order to adjust the system so that it will close out the steady state error, it is sufficient for $A(p)$ to have a simple zero at $p=0$. In this case

$$\omega_1=e^{v_1T_s}\equiv 1, \tag{54}$$

and the only modification of Eq. (24) is that X_1 becomes

$$X_1=(-1)^{m+n-1}W_1W^{-1}T_s. \tag{55}$$

From Eq. (23) for $\omega_1=1$,

$$\sum_{q=0}^{m+n-1}\lambda_q=-1 \quad \text{or} \quad \sum_{q=0}^{m+n}\lambda_q=0 \tag{56}$$

since $\lambda_{m+n}=1$. Hence Eq. (26) reduces to a homogeneous equation and, if the system is stable,

$$\epsilon_N=\sum_{\beta=1}^{m+n}P_\beta r_\beta{}^N\to 0 \quad \text{as} \quad N\to\infty. \tag{57}$$

Now the output θ_N satisfies Eq. (42). Since $\epsilon_N \to 0$ as $N \to \infty$, Eq. (49) can be written

$$\theta_N = \sum_{\alpha=1}^{n} Q_\alpha \sigma_\alpha{}^N + \delta \tag{58}$$

where δ is an infinitesimal for $N \to \infty$. Equation (58) gives the output at the sampling instants for a step input.

The asymptotic value of the output is readily obtained. Since the system is stable by hypothesis and $\phi(t)$ is bounded, $\theta(t)$ must be bounded. This implies

$$|\sigma_\alpha| \leqq 1, \quad \alpha = 1, 2, \cdots, n. \tag{59}$$

But $\sigma_\beta = \omega_\beta$, $\beta = 1, 2, \cdots, n$ and since $\omega_1 = 1$, $\sigma_1 = 1$. All the other σ_α must be less than one since $v_1 = 0$ is the only pole of $M/A(p)$ at $p = 0$. Hence

$$\theta_N = Q_1 + \sum_{\alpha=2}^{n} \sigma_\alpha{}^N + \delta, \tag{60}$$

and thus

$$\theta_\infty = Q_1. \tag{61}$$

It is noted that the condition $|\sigma_\alpha| \leqq 1$ implies Real part $(u_\alpha) \leqq 0$, $\alpha = 1, 2, \cdots, n$. In other words, a necessary condition that the system be stable is that the forward transmission have no poles in the right-half of the complex p plane.

Variation of Elastic Moduli and Wave Velocity with Pressure and Temperature in Plastics*

D. S. Hughes, E. B. Blankenship, and R. L. Mims†
Department of Physics, University of Texas, Austin, Texas
(Received September 6, 1949)

The variations in velocity of dilatational waves in the pressure range 0 to 15,000 lb./in.² and temperature range 30° to 90°C have been measured for samples of polystyrene, Lucite, and polyethylene. The velocity of rotational waves was also measured for polystyrene and Lucite. In polyethylene no trace of a rotational wave could be identified. The elastic moduli and Poisson's ratio are computed over the experimental range.

INTRODUCTION

A QUICK and simple method of measuring the velocity of waves in rods is to measure the time of transmission of a pulse through a known length. This pulse in long thin rods usually travels with the "acoustic" velocity given by

$$V_S = (E/\rho)^{\frac{1}{2}} \tag{1}$$

where E is the Young's modulus of the material and ρ is the density. For short, relatively thick rods the first arrival is the dilatational wave whose velocity is given by

$$V_D = [(K + \tfrac{4}{3}\mu)/\rho]^{\frac{1}{2}} \tag{2}$$

where K is the bulk modulus and μ is the rigidity modulus. It has been found that following the dilatational wave[1] there are other arrivals which arise by transformation of the original dilatational pulse at the boundaries of the rod into rotational waves. These rotational waves cross the rod and at the opposite boundary give rise again to a dilatational wave traveling parallel to the axis of the rod. The angle between the normal to the rotational waves and the axis of the rod is given by

$$\sin\Theta = V_R/V_D. \tag{3}$$

In reference 1 it is shown that the arrival times to be expected are given by

$$t_{nm} = m\alpha D + n\beta L \tag{4}$$

where n is any odd integer and m is subject to the condition

$$0 \leq mD \tan\Theta \leq nL$$

$$\beta = 1/V_D \quad \alpha = [(V_D{}^2 - V_R{}^2)/V_D V_R]^{\frac{1}{2}}. \tag{5}$$

Thus t_{10} is the direct dilatational wave, t_{11} is a dilatational wave that was transformed into a rotational wave and back into a dilatational wave, t_{12} has been transformed twice into a rotational wave, t_{30} is a dilatational wave that has traversed the rod three times without transformation, t_{31} has traversed the rod three times with one transformation, etc.

In practice, values of m greater than five or six are rarely observed even with metals of good acoustic properties. With polystyrene, t_{10}, t_{11}, and t_{30} could always be identified. With Lucite, only t_{10} and t_{11} could be read dependably, and with polyethylene, only t_{10}.

The values of α and β are computed from Eq. (4) and the observed times, and from these, the dilatational and rotational velocities may be derived.

$$V_D = \beta^{-1}$$
$$V_R = (\alpha^2 + \beta^2)^{-\frac{1}{2}}. \tag{6}$$

From these velocities and the density ρ, the elastic constants may be derived.

Define

$$R = V_D/V_R. \tag{7}$$

* This work was supported by the ONR under Contract N6-onr-266, Task Order VIII.
† Employed under research grant from Shell Oil Company.
[1] D. S. Hughes, W. L. Pondrom, and R. L. Mims, Phys. Rev. 75, 1552 (1949).

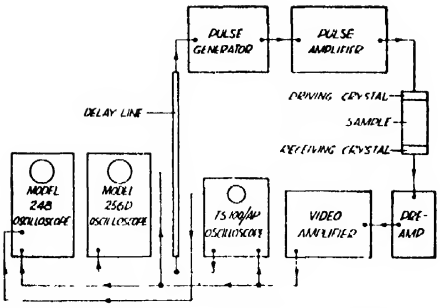


FIG. 1. Block diagram of electronic equipment.

Then, Poisson's ratio is

$$\sigma=(\tfrac{1}{2}R^2-1)/(R^2-1),$$

the rigidity modulus is

$$\mu=\rho V_R^2,$$

the Young's modulus is

$$E=2(1+\sigma)\mu,$$

and the bulk modulus is

$$K=E/3(1-2\sigma).$$

APPARATUS

Figure 1 shows a block diagram of the apparatus. This is essentially the same as described in reference 1. The TS-100/AP oscilloscope is used as standard for all time measurements. This instrument has a circular sweep of 80.86-kc frequency of 12.368 μsec. per revolution. A trigger circuit is provided that operates at submultiples of this frequency in the range 150 to 500 pulses per second. This trigger is accurately synchronized with the circular sweep. The instrument is also provided with an intensity gate so that any one selected revolution, after the trigger, appears on the screen. The trigger is sent through a delay line to synchronize the pulse generator. The pulse generator provides pulses with a rise time of about 0.3 μsec., 180-v amplitude and length from 0.5 to 60 μsec. A volume control was added so that the amplitude could be varied from 20 to 180 v. For materials with poor transmission these pulses are insufficient. A pulse amplifier consisting of a 6AG7 tube to invert and sharpen the pulse, an 807 as an amplifier, and a second 807 as a cathode follower for an output tube was constructed. This amplifier when driven by the pulse generator will provide pulses of about 800-v amplitude and rise time of 0.2 μsec., with an exponential decay of about 5 μsec. This sharp rise and slow decay both impressed on the driving crystal give a very simple pulse in the sample. The driving and receiving crystals are 3.5 Mc, x-cut, quartz crystals, 1.125 in. in diameter.

The signal from the receiving crystals is amplified by a preamplifier similar to that described by Jordan and Bell[2] but modified so that the gain is about 10 and the band width extends to some 5 Mc. This is followed by a video amplifier[3] which has a gain of about 800 and an upper frequency limit of about 8 Mc. The output of this amplifier is sent to the radial deflection electrode of the TS-100/AP and the vertical plates of the 248 and 256D oscilloscopes. The 25-μsec. R sweep of the 256D is triggered by the TS-100/AP generator. This scope is used for qualitative observations of wave form and approximate time checks. The 248 is triggered by the delayed intensity gate of the TS-100/AP so that any delayed 25- or 100-μsec. portion of the record may be displayed. This scope is fitted with a blue tube and camera for photography.

The sample may be used either in open air or placed in the pressure chamber, Fig. 2. The pressure chamber is constructed of Type 309 stainless steel, 4 in. in diameter by 12 in. high. The inner chamber is 2 in. in diameter by 7 in. high. The cylinder sets in a brass tank filled with SAE 30 oil and fitted with heating element, circulating stirrer, and cooling coils. The leads from the crystals are brought out through the head of the cylinder.

DATA

Figure 3 shows records of the waves received through polystyrene, Lucite, and polyethelene at room tem-

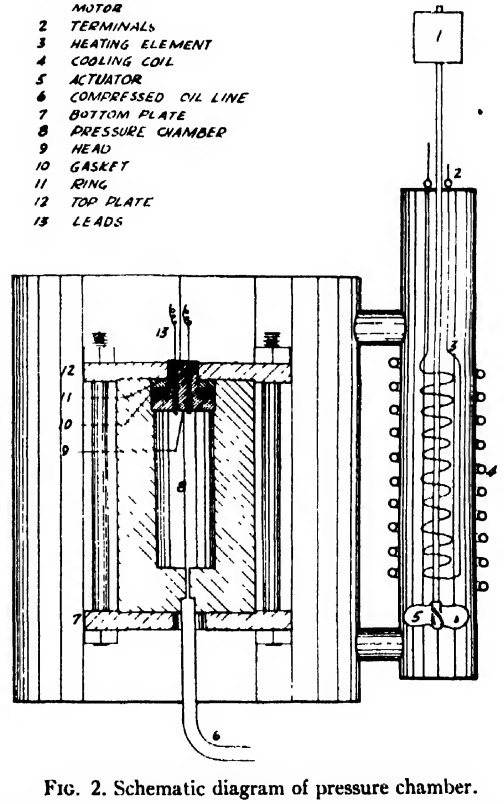


FIG. 2. Schematic diagram of pressure chamber.

[2] W. H. Jordan and P. R. Bell, Rev. Sci. Inst. 18, 703 (1947).
[3] Research Laboratory Series, *Electronic Instruments* (McGraw-Hill Book Company, Inc., New York, 1948), Vol. 21, p. 602.

TABLE I. Material: polystyrene.

Pressure lb./in.²	L=5.090 cm V_D m/sec.	V_R m/sec.	ρ=1.046 g/cc $E\times10^{11}$ dynes/cm²	$\mu\times10^{11}$ dynes/cm²	D=2.626 cm $K\times10^{11}$ dynes/cm²	σ
			Temperature 31°C			
P_0	2295	1139	0.3627	0.1358	0.3699	0.336
2500	2336	1149	0.3703	0.1381	0.3880	0.340
5000	2375	1157	0.3777	0.1404	0.4056	0.342
7500	2413	1169	0.3874	0.1435	0.4233	0.345
10000	2448	1181	0.3952	0.1465	0.4348	0.348
12500	2481	1187	0.4012	0.1484	0.4520	0.351
15000	2515	1194	0.4063	0.1500	0.4598	0.353
			Temperature 51°C			
P_0	2256	1121	0.3505	0.1311	0.3572	0.337
2500	2304	1134	0.3604	0.1345	0.3749	0.341
5000	2344	1147	0.3709	0.1382	0.3976	0.344
7500	2380	1156	0.3773	0.1402	0.4079	0.347
10000	2418	1167	0.3816	0.1427	0.4256	0.349
12500	2449	1174	0.3923	0.1453	0.4382	0.352
15000	2488	1185	0.3989	0.1476	0.4551	0.354
			Temperature 72°C			
P_0	2238	1106	0.3422	0.1278	0.3532	0.339
2500	2275	1118	0.3506	0.1307	0.3675	0.342
5000	2325	1129	0.3586	0.1332	0.3837	0.346
7500	2359	1142	0.3675	0.1364	0.4024	0.348
10000	2399	1153	0.3743	0.1385	0.4151	0.351
12500	2420	1158	0.3815	0.1409	0.4300	0.353
15000	2456	1167	0.3876	0.1428	0.4457	0.355
			Temperature 82°C			
P_0	2205	1090	0.3328	0.1242	0.3390	0.339
2500	2250	1108	0.3444	0.1284	0.3588	0.341
5000	2296	1125	0.3560	0.1321	0.3768	0.347
7500	2339	1136	0.3631	0.1349	0.3945	0.350
10000	2384	1149	0.3723	0.1376	0.4136	0.352
12500	2407	1154	0.3779	0.1398	0.4237	0.354
15000	2434	1163	0.3832	0.1419	0.4352	0.355
			Temperature 92°C			
P_0	2177	1074	0.3237	0.1208	0.3345	0.340
2500	2231	1097	0.3376	0.1258	0.3557	0.344
5000	2277	1118	0.3504	0.1305	0.3703	0.348
7500	2323	1128	0.3578	0.1328	0.3893	0.350
10000	2364	1140	0.3669	0.1359	0.4059	0.351
12500	2398	1149	0.3751	0.1383	0.4217	0.355
15000	2436	1159	0.3809	0.1409	0.4305	0.356

TABLE II. Material: Lucite.

Pressure lb./in.²	L=7.976 cm V_D m/sec.	V_R m/sec.	ρ=1.171 g/cc $\mu\times10^{11}$ dynes/cm²	$E\times10^{11}$ dynes/cm²	D=2.540 cm $K\times10^{11}$ dynes/cm²	σ
			Temperature 33°C			
P_0	2621	1296	0.1968	0.5226	0.5422	0.338
2500	2666	1320	0.2038	0.5453	0.5602	0.338
5000	2702	1336	0.2102	0.5622	0.5768	0.338
7500	2744	1361	0.2175	0.5816	0.5944	0.337
10000	2782	1375	0.2222	0.5949	0.6135	0.338
12500	2820	1390	0.2274	0.6091	0.6324	0.340
15000	2853	1400	0.2310	0.6195	0.6505	0.341
			Temperature 48°C			
P_0	2538	--	—	—	—	--
2500	2585	—	—	—	—	—
5000	2632	1276	0.1912	0.5150	0.5580	0.347
7500	2681	1294	0.1972	0.5312	0.5798	0.347
10000	2719	1315	0.2029	0.5469	0.5978	0.348
12500	2763	1329	0.2075	0.5595	0.6141	0.348
15000	2800	1336	0.2116	0.5719	0.6405	0.351
			Temperature 71°C			
P_0	2468	—	—	—	—	--
2500	2496	—	—	—	—	--
5000	2547	—	—	—	—	--
7500	2620	1249	0.1840	0.4966	0.5505	0.350
10000	2652	1266	0.1889	0.5228	0.5866	0.352
12500	2705	1289	0.1951	0.5386	0.6091	0.353
15000	2738	1305	0.2008	0.5446	0.6308	0.356

perature and pressure. With increase in pressure the arrivals become slightly sharper and, with increase in temperature, somewhat less sharp; but the changes are minor except that with Lucite the t_{11} arrival disappears at low pressures and elevated temperatures. The trace length is about 100 μsec. The arrivals falling in this range are indicated.

The arrivals can be read on the TS-100/AP with a probable error of about ±0.03 μsec. This permits a determination of V_D and V_R with a probable error of about 0.2 percent. From these velocites and the density ρ, Poisson's ratio σ, the rigidity modulus μ, Young's modulus E, and the bulk modulus K, may be computed using Eq. (7). These values are slightly erroneous since the length, diameter, and density as measured at room temperature and pressure are used. However, this first computation gives an approximate value for K as a function of pressure at a given temperature. The length at any given temperature and room pressure may be computed from a knowledge of the coefficient of thermal expansion. Thus,

$$L_{T_0}=L_0[1+\alpha(T-T_0)], \tag{8}$$

where α is the coefficient of thermal expansion, T_0 is room temperature, T is temperature at which the run was made, L_0 is the length or diameter at room temperature and pressure, and L_{T_0} is the computed length at temperature T and room pressure. From the approximate value of K already computed, the length at any temperature and pressure is given by

$$L_{T_P}=L_{T_0}[1-(P/3\bar{K})], \tag{9}$$

where P is the pressure and $\bar{K}$ is average bulk modulus over the pressure range zero to P. Using Eqs. (8) and (9), the length and diameter were computed for all pressures and temperatures. The density ρ and the velocities V_D and V_R were recomputed as a function of pressure and temperature and then Eq. (8) used to re-

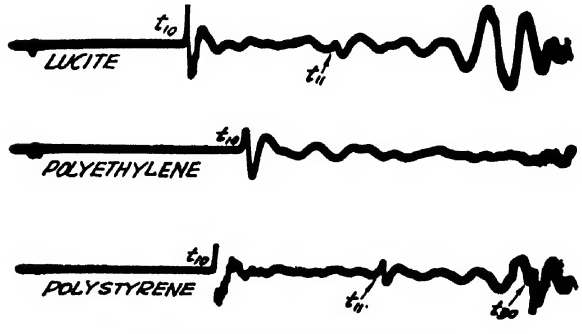


FIG. 3. Pulse transmission in plastics.

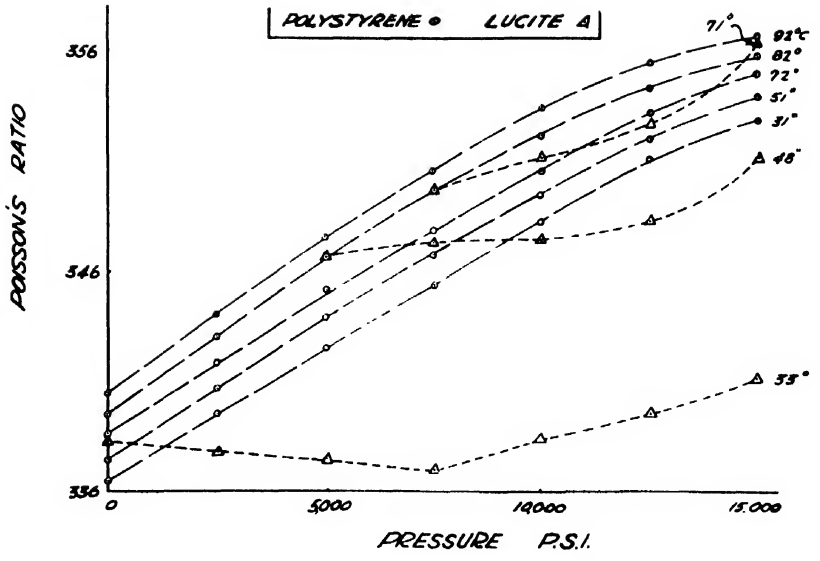


FIG. 4. Comparison of Poisson's ratio for Lucite and polystyrene.

compute values of σ, μ, E, and K. The largest correction is about two percent. It is assumed that for the purpose of this correction the differences between the adiabatic and isothermal moduli can be neglected.

Table I lists the velocities and elastic moduli for polystyrene over the range of pressures and temperatures. The variation in the rotational velocity with both temperature and pressure is very much less than that in the dilatational velocity. Similarly, the rigidity modulus increases by approximately six percent for an increase in pressure of 15,000 p.s.i. whereas the bulk modulus increases over the same range by about 30 percent. Corresponding to this rise in bulk modulus, Poisson's ratio, σ, increases uniformly.

Table II lists the velocities and elastic moduli for Lucite. Although the elastic moduli vary in a somewhat similar way to those of polystyrene there are fundamental differences. At room temperature Poisson's ratio decreases slightly as the pressure is raised reaching a minimum at 7500 p.s.i. and then increases slightly thereafter. At 48°C the first transformed wave cannot be read until a pressure of 5000 p.s.i. is reached and at 71°C a pressure of 7500 p.s.i. is required.

This almost certainly indicates a large change in viscous damping of the rotational waves. The first appearance of the transformed waves as the pressure increases is not sudden but gradual. The pressures listed above are those at which a clear definite arrival can be identified. At the elevated temperatures, Poisson's ratio rises very rapidly with pressure in contrast to its behavior at room temperature for Lucite and polystyrene over the experimental range as shown by Fig. 4.

Table III lists the dilatational velocity for polyethylene. No trace of a transformed wave could be detected and the elastic constants could not be computed.

TABLE III. Material: polyethylene.

	l = 5.093 cm		ρ = 0.9069 g/cc		D = 2.535 cm	
Pressure lb/in.²			V_D m/sec.			
Temp	31°C	45°C	51°C	55°C	68°C	73°C
P_0	1979	1844	1807	1725	1678	1643
2500	2064	1918	1876	1837	1753	1716
5000	2129	1991	1939	1904	1829	1792
7500	2193	2061	2010	1973	1897	1864
10000	2257	2123	2079	2034	1961	1925
12500	2320	2179	2137	2095	2027	1986
15000	2374	2227	2191	2152	2082	2040

The Summation of Fourier Series by Operational Methods

LOUIS A. PIPES
Department of Engineering, University of California, Los Angeles, California
(Received September 26, 1949)

A method for the summation of certain types of Fourier series by the use of Laplace transforms is presented. The procedure enables one to easily obtain graphical representations of functions defined by Fourier series in many cases of practical importance.

I. INTRODUCTION

IN the mathematical analysis of problems in electric circuit theory, mechanical vibrations, heat conduction, and other branches of mathematical physics, one frequently arrives at solutions that are expressed in the form of infinite Fourier series. In many cases, it is more convenient to have the solutions of these problems expressed in closed form for purposes of calculation.

Oliver Heaviside[1] in the second volume of his *Electromagnetic Theory* used his rather informal "resistance operator" methods to interpret and sum Fourier series. As is well known, his mathematical processes were sometimes open to question. However, his methods frequently lead to useful results most simply, and he himself said in several instances, "I feel quite certain that this could be shown by conventional mathematics."

It is the purpose of this discussion to present a method for summing certain types of Fourier series of frequent occurrence in applied mathematics by the use of the modern Laplace method of the "Operational Calculus."

II. FUNDAMENTAL TRANSFORMATIONS

The method to be described is based upon the fact that certain hyperbolic functions can be expressed as an infinite series of exponential functions. If we use the notation,

$$A_k = e^{-kap} = \exp(-kap) \tag{2.1}$$

then the following expansions may be written

$$\tanh(ap) = 1 - 2A_2 + 2A_4 - 2A_6 + \ldots, \tag{2.2}$$

$$\coth(ap) = 1 + 2A_2 + 2A_4 + 2A_6 + \ldots, \tag{2.3}$$

$$\frac{1}{\sinh(ap)} = 2(A_1 + A_3 + A_5 + A_7 + \cdots) \tag{2.4}$$

$$\frac{1}{\cosh(ap)} = 2(A_1 - A_3 + A_5 - A_7 + \cdots), \tag{2.5}$$

a is a positive real number, and p is a number whose real part is positive.

By the use of a method due to Mittag-Leffler,[2] it is possible to obtain the following expansions,

$$\tanh(ap) = \frac{2}{ap}\left\{\frac{1}{1+(\pi/2ap)^2} + \frac{1}{1+(3\pi/2ap)^2} + \frac{1}{1+(5\pi/2ap)^2} + \cdots\right\} = \frac{2}{ap}\sum_{n=1}^{\infty}\frac{1}{1+\{[(2n-1)\pi]/2ap\}^2}, \tag{2.6}$$

$$\coth(ap) = \frac{1}{ap}\left\{1 + \frac{1}{1+(\pi/ap)^2} + \frac{2}{1+(2\pi/ap)^2} + \cdots\right\} = \frac{1}{ap}\left\{1 + \sum_{n=1}^{\infty}\frac{2}{1+(n\pi/ap)^2}\right\}, \tag{2.7}$$

$$\frac{1}{\cosh(ap)} = \frac{1}{a^2p^2}\left\{\frac{\pi}{1+(\pi/2ap)^2} - \frac{3\pi}{1+(3\pi/2ap)^2} + \cdots\right\} = \frac{1}{a^2p^2}\sum_{n=1}^{\infty}\frac{(-1)(2n-1)\pi}{1+(n\pi/2ap)^2}, \tag{2.8}$$

$$\frac{1}{\sinh(ap)} = \frac{1}{ap}\left\{1 - \frac{2}{1+(\pi/ap)^2} + \frac{2}{1+(2\pi/ap)^2} - \cdots\right\} = \frac{1}{ap}\left\{1 + \sum_{n=1}^{\infty}(-1)^n\frac{2}{1+(n\pi/ap)^2}\right\}. \tag{2.9}$$

It is possible to expand other combinations of hyperbolic functions in a similar manner, however, the above expansions are sufficient to illustrate the general principles of the method to be described.

III. THE LAPLACE TRANSFORMATION

Modern rigorous "operational calculus" is based upon the Fourier-Mellin integral equations.[3]

[1] O. Heaviside, *Electromagnetic Theory* (D. Van Nostrand Company, Inc., New York, 1925), Vol. II, pp. 107–119.
[2] E. T. Whittaker, and G. N. Watson, *Modern Analysis* (Cambridge University Press, London, 1927), pp. 134–136.
[3] L. A. Pipes, *Applied Mathematics for Engineers and Physicists* (McGraw-Hill Book Company, Inc., New York, 1946), Chapter XXI.

These equations state that, if

$$g(p)=p\int_0^\infty e^{-pt}h(t)dt, \tag{3.1}$$

then

$$h(t)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}\frac{g(p)e^{pt}dp}{p}, \tag{3.2}$$

where $j=(-1)^{\frac{1}{2}}$ and c is a real constant of such magnitude that the poles of the integrand lie entirely to the left of the path of integration. For brevity, the notation

$$g(p)=Lh(t) \quad \text{or} \quad h(t)=L^{-1}g(p) \tag{3.3}$$

is used to denote the above relations between the functions $g(p)$ and $h(t)$. The function $g(p)$ is said to be the Laplace transform of $h(t)$ and the function $h(t)$ is said to be the inverse Laplace transform of $g(p)$. Extensive tables of Laplace transforms have been compiled.[3] In this discussion we require the transforms of the sine and the cosine. These transforms are given by

$$L\sin(nwt)=\frac{nwp}{p^2+n^2w^2}=\frac{nw}{p}\cdot\frac{1}{1+(nw/p)^2} \tag{3.4}$$

and

$$L\cos(nwt)=\frac{p^2}{p^2+(nw)^2}=\frac{1}{1+(nw/p)^2}. \tag{3.5}$$

Another result of the theory of Laplace transform that is required is the following one.[3]

If

$$L^{-1}g(p)=h(t) \tag{3.6}$$

then

$$L^{-1}e^{-kp}g(p)=\begin{cases}0 & t<k\\ h(t-k) & t>k,\end{cases} \tag{3.7}$$

where k is a positive constant.

IV. SUMMATION OF FOURIER SERIES

The general method of summation will now be illustrated by applying it to several special cases.

Example I

In certain investigations of physical problems the following Fourier series is obtained:

$$F_1(t)=4[(\sin t/1)+(\sin 3t/3)+(\sin 5t/5)+\cdots]$$

$$=4\sum_{n=1}^{\infty}\frac{\sin(2n-1)t}{(2n-1)} \qquad t>0. \tag{4.1}$$

The Laplace transform of the function $F(t)$ represented by the series (4.1) may be obtained by taking the Laplace transform of each term of the series using (3.4). In this manner one obtains

$$LF_1(t)=\frac{4}{p}\left[\frac{1}{1+(1/p)^2}+\frac{1}{1+(3/p)^2}+\frac{1}{1+(5/p)^2}+\cdots\right]. \tag{4.2}$$

We now note that, if in Eq. (2.6) we let

$$a=\pi/2, \tag{4.3}$$

then we can express the series term of (4.2) in the following form:

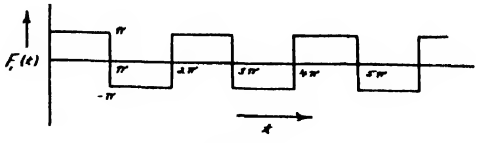


FIG. 1.

$$\left[\frac{1}{1+(1/p)^2}+\frac{1}{1+(3/p)^2}+\frac{1}{1+(5/p)^2}+\cdots\right]$$

$$=ap/2\tanh(ap). \tag{4.4}$$

Hence, comparing (4.2) and (4.4) we see that

$$LF_1(t)=2a\tanh(ap)=\tanh(\pi p/2). \tag{4.5}$$

We now use the expansion (2.2) and obtain

$$LF_1(t)=\pi[1-2e^{-\pi p}+2e^{-2\pi p}-2e^{-3\pi p}+\ \ldots]. \tag{4.6}$$

This is the Laplace transform of the Fourier series (4.1). To obtain the inverse Laplace transform of (4.6) and then obtain a representation of the series (4.1) we use the result (3.7). The inverse Laplace transform of the first term of (4.6) is a step function of height π formed at $t=0$. The inverse transform of the second term is a step function of height -2π formed at $t=\pi$. The inverse transform of the third term is a step function of height $+2\pi$ formed at $t=2$ and so on. The sum of all these step functions is illustrated in Fig. 1.

Example II

As a second example let it be required to interpret the following Fourier series that occurs in certain electrical and heat conduction problems.

$$F_2(t)=\frac{2}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi t}{s}\right). \tag{4.7}$$

By the use of (3.4), the Laplace transform of (4.7) is obtained and is given by:

$$LF_2(t)=\frac{2}{\pi}\sum_{n=1}^{\infty}\frac{n\pi}{nsp}\cdot\frac{1}{1+(n\pi/sp)^2}$$

$$=\frac{1}{sp}\sum_{n=1}^{\infty}\frac{2}{1+(n\pi/sp)^2}. \tag{4.8}$$

If we now compare (4.8) with (2.7) it is seen that we have,

$$LF_2(t)=\coth(sp)-1/sp. \tag{4.9}$$

By the use of (2.3) this may be written in the following form,

$$LF_2(t)=(1-1/sp+2e^{-2sp}+2e^{-4sp}+2e^{-6sp}+\ldots). \tag{4.10}$$

To obtain the inverse transform of F_2 we note that the inverse transform of the first term is a unit step function, the inverse transform of the term $-1/sp$ is $-t/s$, the inverse transform of $+2e^{-2sp}$ is a step function of height that begins at $t=2s$, etc. The sum of the various transforms gives the function $F_2(t)$. The result is given graphically in Fig. 2.
This figure is a graphical representation of the Fourier series $F_2(t)$.

Example III

As another example, let it be required to obtain the graph of the following Fourier series:

$$F_3(t)=(\sin t-\tfrac{1}{2}\sin 2t+\tfrac{1}{3}\sin 3t-\cdots)$$

$$=-\sum_{n=1}^{\infty}(-1)^n\frac{1}{n}\sin(nt). \tag{4.11}$$

In this case we transform (4.11) by the use of Eq. (3.4) and note the similarity to Eq. (2.8). Then the expansion (2.4) gives the result

$$LF_3(t)=(1/p-2\pi e^{-\pi p}-2\pi e^{-3\pi p}-2\pi e^{-5\pi p}+\ldots). \tag{4.12}$$

Now the inverse transform of $1/p$ is t, the inverse transform of $-2\pi e^{-\pi p}$ is a step function of height -2π that begins at $t=\pi$, etc. The sum of the various inverse transforms is $F_3(t)$. This function is given graphically in Fig. 3.

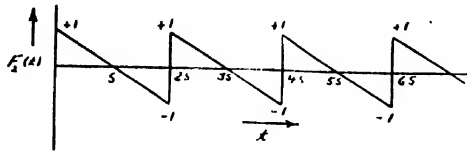


Fig. 2.

Example IV

An interesting series which arises in the analysis of the propagation of charge along a cable will now be considered. This series represents a row of impulses and the function it represents is important in the analysis of the disturbances created when lightning hits a transmission line. The series to be considered is the following one:

$$F_4(x)=\frac{2}{s}\sum_{n=1}^{n=\infty}\sin\left(\frac{n\pi x}{s}\right)\sin\left(\frac{n\pi b}{s}\right). \tag{4.13}$$

To study this series more easily, the product of sines may be transformed into two cosines and written in the following form:

$$F_4(x)=\frac{1}{s}\sum_{n=1}^{n=\infty}\left[\cos\frac{n\pi(x-b)}{s}-\cos\frac{n\pi(x+b)}{s}\right]. \tag{4.14}$$

Let us first consider the first series of (4.14) and let,

$$t=x-b. \tag{4.15}$$

This series may be written in the form:

$$\phi(t)=\frac{1}{s}\sum_{n=1}^{\infty}\cos\left(\frac{n\pi st}{s}\right). \tag{4.16}$$

By (3.5) it can be transformed to:

$$L\phi(t)=\frac{1}{s}\sum_{n=1}^{\infty}\cos\frac{1}{1+(n/sp)}. \tag{4.17}$$

By the use of Eqs. (2.7) and (2.3) this may be written in the form,

$$L\phi(t)=p/s(1+2e^{-2sp}+2e^{-4sp}+\ldots). \tag{4.18}$$

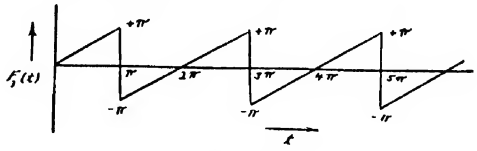


Fig. 3.

To obtain the inverse transform of this function we note that,[3]

$$L^{-1}p=\delta(t). \tag{4.19}$$

The function, $\delta(t)$, is an impulse function known in the literature as the Dirac function. It has the following properties:

$$\delta(t)=\begin{cases}0 & t\neq 0\\ \infty & t=0\end{cases} \tag{4.20}$$

and

$$\int_{-\infty}^{+\infty}\delta(t)dt=1. \tag{4.21}$$

As a consequence of (3.7), the inverse transform of (4.18) is

$$\phi(t)=1/s[\delta(t)+2\delta(t-2s)+2\delta(t-4s)+\ldots]. \tag{4.22}$$

We note that the first term is an impulse at $t=x-b=0$ or at $x=b$. The second term is an impulse located at $t=2s$ or $x=2s+b$, and so forth. The second series of (4.14) may be transformed in a similar manner if the substitution $t=x+b$ is made and it can be shown that this series represents a row of negative impulses at $x=-b$, $x=2s-b$, etc. A graphical representation of the function $F_4(x)$ defined by (4.13) is given in Fig. 4.

The function $F_4(x)$ is zero for all values of x except at the places where it has the values of positive and negative pulses of the Dirac type.

V. GENERAL CONSIDERATIONS

The general Fourier series may be written in the form:

$$F(t)=A_0+\sum_{n=1}^{\infty} A_n \cos(nwt)+\sum_{n=1}^{\infty} B_n \sin(nwt). \quad (5.1)$$

The coefficients A_n and B_n are known functions of n. By the use of Eqs. (3.4) and (3.5), it is possible to transform (5.1) into

$$LF(t)=A_0+\sum_{n=1}^{\infty} \frac{An}{1+(nw/p)^2}+\frac{w}{p}\sum_{n=1}^{\infty} \frac{nBn}{1+(nw/p)^2}. \quad (5.2)$$

The success of the method of interpretation described in this discussion depends upon the possibility of effecting the summations of (5.2) by the use of the Eqs. (2.6)–(2.9) or by formulas derivable from them.

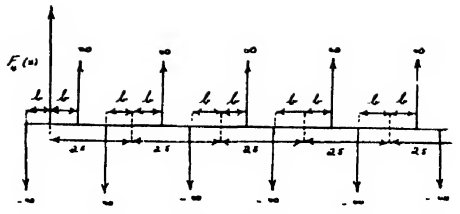


FIG. 4.

The interpretation of the Fourier series can then be carried out by the use of Eqs. (2.2)–(2.5) or combinations of these expansions.

In many cases, however, the summation of the fractions of (5.2) may present insurmountable difficulties and in such cases, this method of interpretation is not applicable.

Effect of Change of Scale on Sintering Phenomena

CONYERS HERRING
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received November 16, 1949)

It is shown that when certain plausible assumptions are fulfilled simple scaling laws govern the times required to produce, by sintering at a given temperature, geometrically similar changes in two or more systems of solid particles which are identical geometrically except for a difference of scale. It is suggested that experimental studies of the effect of such a change of scale may prove valuable in identifying the predominant mechanism responsible for sintering under any particular set of conditions, and may also help to decide certain fundamental questions in fields such as creep and crystal growth

FOR the purposes of this paper "sintering" will be understood to mean any change in shape which a small particle or a cluster of particles of uniform composition undergoes when held at high temperature. In the absence of externally applied stresses, such changes are believed to be motivated by surface tension, since the surface free energy decreases as the particles grow together and assume a more compact shape. The transport of matter which is involved can take place by any of four mechanisms: viscous or plastic flow, evaporation and condensation, volume diffusion, and surface migration. With the aid of certain plausible assumptions the rates of transport by each of these mechanisms can be computed in terms of the geometrical parameters and certain constants of the material such as surface tensions, diffusion coefficients, etc.[1] Foremost among the assumptions involved is the assumption that the building up or tearing down of the surface can occur quite readily in response to conditions in the immediately adjoining solid or vapor phase. This assumption is closely related to the controversial problem of the mechanism of crystal growth, and it would be incorrect, at least in cases involving plane crystal facets, if growth normal to such facets required a "two-dimensional nucleation" process.[1,2] It is the purpose of this note to point out that useful information can probably be obtained, both on the fundamental question of the way in which crystals grow and on the more specific question of the relative importance of the four transport mechanisms in specific cases, by a comparison of sintering rates for particles or clusters of different sizes but geometrically similar shapes.

THE SCALING LAWS

We shall use the term "cluster" to refer to any particle or to a number of particles which have started to grow together. Consider two partially sintered clusters which are geometrically similar, the linear dimensions of number two being λ times those of number one. For example, if each cluster consists of spherical

[1] C. Herring, paper presented at the Symposium on the Physics of Powder Metallurgy organized by Sylvania Electric Products, Inc., Bayside, Long Island, New York (August 24–26, 1949). Plans have been made to publish the papers of the Symposium in book form.

[2] The theory of growth by two-dimensional nucleation is summarized in the book of M. Volmer, *Kinetik der Phasenbildung* (Verlag von Theodor Steinkopff, Dresden and Leipzig, 1939; Edwards Brothers, Inc., Ann Arbor, 1945). An alternative theory of crystal growth has been proposed by F. C. Frank, to appear in *Trans. Faraday Soc.* (paper presented at the Discussion on Crystal Growth sponsored by the Faraday Society and held at Bristol, England, April, 1949).

grains of radius R, sintered until the radius of the area of contact of adjacent grains is a, similarity will require $R_2=\lambda R_1$, $a_2=\lambda a_1$. The question to be considered is, how is the time Δt_1 required to produce a given change in shape of cluster number one in the sintering process related to the time Δt_2 required, at the same temperature, to produce a geometrically similar change in the second cluster? This relation turns out to be quite simple in the following four cases, if the same one of the basic mechanisms predominates for both clusters and if the above-mentioned assumption of facile growth and recession of the surface is valid.

Viscous Flow of an Amorphous Material

It is well known that the surface tension γ of a spherical drop of liquid of radius R produces a pressure p in the interior of the drop, given by

$$2\pi R\gamma=\pi R^2 p$$

or

$$p=2\gamma/R. \tag{1}$$

This equation can be shown to give, in more general cases, the normal stress p_{rr} immediately under a curved surface of any shape, provided R^{-1} is interpreted as the mean of the principal curvatures of the surface, these being counted positive if convex, negative if concave. It can be applied to "amorphous solids" provided the temperature is high enough so that the surface layers of atoms are reasonably free to rearrange themselves; however, the simple form (1) is not valid for crystals.[1]

If the surface of a cluster is inhomogeneously curved, the normal stresses given by (1) will vary over the surface, and to compensate this there must be shearing stresses in the interior. For the two sizes of cluster mentioned, the stress tensors at corresponding points, being determined by (1) and the linear conditions of equilibrium, must satisfy

$$p_{ij}^{(2)}(\lambda\mathbf{r})=p_{ij}^{(1)}(\mathbf{r})/\lambda. \tag{2}$$

Thus assuming viscous behavior the strain rates ds_{ij}/dt at corresponding points will satisfy

$$(ds_{ij}/dt)_2=(1/\lambda)(ds_{ij}/dt)_1$$

and so the times required for a given change of shape to take place will be related by

$$\Delta t_2=\lambda\Delta t_1. \tag{3}$$

Evaporation and Condensation

It will be assumed that the mean free path of an atom or molecule in the vapor is long compared with the dimensions of the particles being sintered, and that only one kind of atom or molecule is involved. For a liquid or an amorphous solid it is well known that the equilibrium vapor pressure over a spherically curved surface of radius R is greater than that over a flat surface by the ratio[3]

$$p_R/p_\infty=\exp(2\Omega_0\gamma/kTR), \tag{4}$$

where Ω_0 is the atomic or molecular volume and γ the surface tension. Application of a similar thermodynamic argument to crystals yields a relation which, though more complicated, agrees with (4) in having an exponent which varies inversely with the first power of the linear scale.[1] If the exponent is small, as is likely to be the case if R is of the order of 10^{-5} cm or greater, (4) can be written

$$p_R=p_\infty(1+2\Omega_0\gamma/kTR). \tag{5}$$

Now consider any two elements of surface dS_a and dS_b on nearby portions of the cluster being sintered. The evaporation rates from these two will be in the ratio of their areas and of their equilibrium vapor pressures.* If these vapor pressures are different, as they will be if the radii of curvature R_a and R_b are different, the amount of material which in any given time is evaporated from a, passes directly to b, and condenses there, will be different from the amount passing in the reverse direction. Ignoring factors of proportionality which are independent of the linear scale factor λ, one may write

$$\text{rate } (a\to b)-\text{rate } (b\to a)\propto dS_a dS_b(p_a-p_b)/r_{ab}^2$$

where r_{ab} is the distance of dS_a from dS_b. Thus the rates for the two geometrically similar aggregates we are considering bear the ratio

$$\frac{\text{rate } (a\to b)_2-\text{rate } (b\to a)_2}{\text{rate } (a\to b)_1-\text{rate } (b\to a)_1}=\lambda^2\cdot\lambda^2\cdot\lambda^{-1}\cdot\lambda^{-2}=\lambda. \tag{6}$$

A similar equation obviously holds for material transported from a to b by any type of trajectory involving reflection from some intermediate surface. Since the total amount of material which must be transported in these ways to produce geometrically similar changes in cluster 2 is λ^3 times that for cluster 1, the intervals of time required will be related by

$$\Delta t_2=\Delta t_1\lambda^3/\lambda=\lambda^2\Delta t_1. \tag{7}$$

Volume Diffusion

It can be shown that in a crystal the diffusive flux of matter, i.e., of interstitial atoms or lattice vacancies, is proportional to the local gradient of the chemical potential of the diffusing substance. (Use of the chemical potential takes account of the effects of both pressure gradients and concentration gradients.) Now the value of this chemical potential at any place on the boundary

[3] See, for example, E. A. Guggenheim, *Modern Thermodynamics by The Methods of Willard Gibbs* (Methuen and Company, Ltd., London, 1933), p. 172.

* This statement may not be true if the "sticking coefficients" α_a and α_b for the two surface elements are different. However, since both the probability of evaporation and the probability of condensation contain the sticking coefficient as a factor, (6) will be correct in any case.

of the crystal must coincide with the chemical potential of the vapor which is in equilibrium with that portion of the crystal surface. For a one-component substance we have from (4), or from its extension to the crystalline case[1]

$$\text{chemical potential at surface} = \text{const.} + kT \ln(p_R/p_\infty) = \text{const.} + B/(\text{linear dimension}), \quad (7)$$

where B is a quantity which is the same at corresponding points of the surfaces of the geometrically similar clusters 1 and 2. Thus, when a quasi-steady distribution of diffusion currents has been set up we have at corresponding points of the interiors of the two clusters

$$(\text{chem. pot.})_2 = (\text{chem. pot.})_1/\lambda$$

and

$$(\text{flux})_2 = (\text{flux})_1/\lambda. \quad (8)$$

Since corresponding areas in the two clusters bear the ratio λ^2 and the amount of material which must be transported goes as λ^3, the times required to produce corresponding amounts of sintering satisfy

$$\Delta t_2 = (\lambda^3 \cdot \lambda^2/\lambda^2)\Delta t_1 = \lambda^3 \Delta t_1. \quad (9)$$

Surface Migration

The scaling law for sintering due to surface migration can be derived in a similar way, by starting from the assumption that the rate of migration over a given type of surface is proportional to the gradient of the chemical potential in the plane of the surface. Thus (8) will hold for the surface flux as well as for the volume flux. In the present case, however, the total rate of transport is the product of the flux not by an area but by a length in the plane of the surface and normal to the flux direction. Thus, instead of (9) we have

$$\Delta t_2 = (\lambda^3 \cdot \lambda^2/\lambda)\Delta t_1 = \lambda^4 \Delta t_1. \quad (10)$$

DISCUSSION

In attempting to draw conclusions from experimental comparisons of sintering times for similar systems of different sizes, several obvious precautions must be observed. One must be sure that the factors determining the degree of contamination of the surface are kept the same for all the cases being compared. If possible, the orientations of the various crystals making up each cluster should be the same for the clusters of different sizes which are being compared: surface tension, surface mobilities, evaporation rates, etc., can be quite different for differently oriented surfaces, the more so because preferential adsorption can cause different surfaces to acquire widely different degrees of contamination. If there is an appreciable delay in the formation of the initial bond between two particles, time should be measured from the moment when this bond is first established.

When experimental data fail to follow one of the scaling laws one must decide whether this is merely because more than one of the above mechanisms is operative over the range of sizes covered by the experiment or whether mechanisms basically different from those assumed above are involved. In the former case there will usually be certain accessible ranges of temperatures and particle sizes over which only one of the four mechanisms contributes appreciably to the changes involved, and if data from these ranges are available, the transition ranges can easily be identified as such. If basically different mechanisms are involved, on the other hand, the variation of sintering time with linear scale may be of a sort quite inconsistent with the hypothesis of a gradual transition from one of the above power laws to another. For example, if a change were produced by slip or creep which set in abruptly at a given critical stress, the sintering time might vary discontinuously with linear scale. An almost discontinuous variation might also result if growth of crystal facets by two-dimensional nucleation were involved.

If a given type of sintering experiment is found to obey one of the scaling laws, the predominant mechanism of transport can be inferred. This inference can probably be drawn with less ambiguity than in the method used by Kuczynski,[4] who inferred the mechanism of sintering by comparing the observed rate of growth of the neck between a particle and a flat block with rates computed theoretically according to each of the above mechanisms. According to Kuczynski's calculations, the radius of the neck varies as the nth root of the sintering time, where $n=2$ for viscous flow, 3 for evaporation and condensation, 5 for volume diffusion, and 7 for surface migration. Although these exponents cannot be far from the truth if the assumptions used in the present paper are correct, uncertainties in geometrical details of the calculations could perhaps affect the values of n by as much as 0.5. The exponents in the present scaling laws, on the other hand, are subject to no such uncertainty.

[4] G. C. Kuczynski, J. Met. 1, 169 (1949).

A Resonant Bar Method for Determining the Elastic Properties of Thin Lamina

A. W. Nolle and P. J. Westervelt
Acoustics Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received September 9, 1949)

A resonant bar method for obtaining the dynamic elastic properties of a thin cement layer is discussed. Formulas are deduced for the calculation of Young's modulus from the measured resonant frequency and band width of two similar steel bars, cemented end-to-end. The accuracy of the method is discussed.

I. INTRODUCTION

THE method to be described has proved useful in obtaining the effective elastic properties of a thin layer of cement. In this application two cylindrical steel bars (see Fig. 1) of equal length $l/2$ are cemented end-to-end forming a single bar. By suitable means the frequency and Q of the fundamental resonance of the cemented bar are measured and compared with similar quantities measured for a continuous bar of length l. From this comparison it is possible to obtain an effective complex Young's modulus for the thin cement film.

For a discussion of the concept of a complex modulus the reader is referred to a recent paper by Nolle.[1] Briefly, the complex modulus is defined as the stress-to-strain ratio for sinusoidal displacements. Provided the displacements are proportional to $e^{-i\omega t}$, the complex modulus E can be expressed as $E=E_1-iE_2$ where E_1 is the ordinary Young's elastic modulus, and $E_2=\omega\gamma$ (γ=normal viscosity coefficient).

The moduli are termed "effective" for two reasons. In the first place, E_1 and γ are presumed to vary with frequency, whereas for a true modulus, E_1 and γ would be independent of frequency. Secondly, in laterally constrained lamina, the measured modulus might lie between that of a Young's modulus and a bulk modulus.

Approximate formulae will first be derived for E_1 and E_2 under the assumption that the resonant frequency of the cemented bar does not differ much from the value for the continuous bar.

II. APPROXIMATE FORMULA FOR E_1

The fundamental resonant frequency ν_0 of a bar of length l is given by

$$l\nu_0=c/2, \tag{1}$$

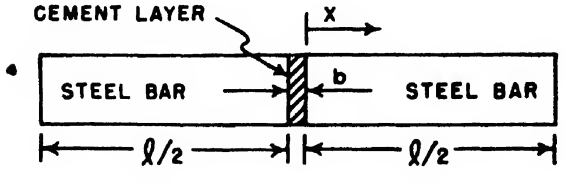


FIG. 1. Sketch of cemented bars.

where the velocity of sound $c=(E/\rho)^{\frac{1}{2}}$ (E=Young's modulus; ρ=density for the bar).

Increasing the length of the bar by δl will result in a decrease $\delta\nu$ of the resonant frequency, given by:

$$\delta\nu/\nu_0=\delta l/l, \tag{2}$$

as can be verified by differentiating Eq. (1). Next, suppose the element $\delta l \ll l$ lies in the middle of the rod. The effect of δl can be visualized as that of a spring, coupling the two halves of the rod together. If the element had a Young's modulus E_1 instead of the modulus E of the bar, its length would have to be $(E_1/E)\delta l$ in order to give the same fractional change in frequency; thus the fractional change due to an element b with modulus E_1 is:

$$\delta\nu/\nu_0=(E/E_1)(b/l). \tag{3}$$

From Eqs. (1), (3) and the expression $E=\rho c^2$, we obtain an expression for the real part of the complex Young's modulus, E_1, of a lamina of thickness b:

$$E_1=4\rho l b\nu_0^2(\nu_0/\delta\nu). \tag{4}$$

It can be shown that Eq. (4) is valid to within 10 percent provided the band width $\Delta\nu$ of the resonance of the cemented bar is small enough to satisfy $(\Delta\nu)^2 \ll(\delta\nu)^2$, and provided $(\delta\nu/\nu_0)<0.1$. These restrictions are met in many cases. Formulas that hold when the first restriction does not apply will be given in the Appendix.

In the event that $\delta\nu/\nu_0$ is not small with respect to unity but that $(\Delta\nu)^2\ll(\delta\nu)^2$ and $b\ll l$, it can be shown that a more precise value of E_1 is obtained by multiplying the value given by Eq. (4) by the function $F(\delta\nu/\nu_0)$:

$$F\left(\frac{\delta\nu}{\nu_0}\right)=\frac{\pi}{2}\left(1-\frac{\delta\nu}{\nu_0}\right)\left(\frac{\delta\nu}{\nu_0}\right)\cot\frac{\pi\delta\nu}{\nu_0}. \tag{5}$$

From Eq. (5), $F(0.1)=0.90$. Hence, when $\delta\nu/\nu_0=0.1$, the value of E_1 obtained from Eq. (4) will be too high by 10 percent.

Another approximate form results from expanding the cotangent in Eq. (5) and retaining the first term in the expansion. This leads to the following expression which is accurate to within 10 percent, provided $\delta\nu/\nu_0<0.3$:

$$E_1=4\rho l b\nu_0^2\left(\frac{\nu_0}{\delta\nu}-1\right), \tag{6}$$

[1] A. W. Nolle, "Method for measuring dynamic mechanical properties of rubber-like materials," J. App. Phys. 19, 753 (1948).

which may be written:

$$E_1 = 4\rho l b \nu_0^2 (\nu/\delta\nu), \tag{7}$$

where $\nu = \nu_0 - \delta\nu$, the resonant frequency of the cemented bars.

III. APPROXIMATE FORMULA FOR E_2

We shall look for a relation between E_2 (which is a measure of the damping of the lamina), the resonant frequency of the cemented bar ν and the band width $\Delta\nu$ of the resonance peak. The quantity E_2 can be expressed in terms of the mechanical resistance, R, of the lamina which is supposed to have a thickness $b \ll l$ and the same area S as that of the bar:

$$E_2 = (2\pi\nu R b)/S. \tag{8}$$

Provided that the Q of the fundamental mode is at all times large compared with unity, the band width $\Delta\nu$ is given by:[2]

$$\Delta\nu = \frac{\nu}{2\pi} \times \frac{\text{Decrease of Energy of Free Vibration per Cycle}}{\text{Total Energy of Vibration}}. \tag{9}$$

The total energy of vibration in the bar is the maximum kinetic energy, or:

$$\text{total energy} = (MV_0^2)/2, \tag{10}$$

where M is the equivalent mass of the bar, vibrating in its first mode

$$M = (\rho l S)/2, \tag{11}$$

and V_0 is the amplitude of the velocity configuration $V(x)$ at a point $x < l/2$ along the bar, measured from the plane joining the bar and the sample:

$$V(x) = V_0 \cos\left[\left(1 - \frac{\delta\nu}{\nu_0}\right)\left(\frac{\pi x}{l} - \frac{\pi}{2}\right)\right].$$

The energy loss per cycle in the lamina is given by:

$$\text{loss per cycle} = \frac{1}{2}\frac{R 2 V(x - l/2)^2}{\nu},$$

or, for $\delta\nu/\nu_0 \ll 1$:

$$\text{loss per cycle} = \pi^2 V_0^2 \frac{R}{2\nu_0}\left(\frac{\delta\nu}{\nu_0}\right)^2. \tag{12}$$

Introducing Eqs. (10) and (12) into (9) we obtain:

$$\Delta\nu = \frac{\pi}{2}\frac{R}{M}\left(\frac{\delta\nu}{\nu_0}\right)^2,$$

then, solving for R and using Eq. (11)

$$R = \frac{\rho l S}{\pi}\left(\frac{\nu_0}{\delta\nu}\right)^2 \Delta\nu.$$

Introducing the above expression into Eq. (8):

$$E_2 = 2\rho l b \nu_0 (\nu_0/\delta\nu)^2 \Delta\nu,$$

or in terms of E_1, given in Eq. (4):

$$E_2 = (E_1/2)\Delta\nu/\delta\nu. \tag{13}$$

Any loss inherent in the continuous unloaded bar is effectively added to the loss introduced by the sample. Therefore, if the band width of the unloaded rod is not negligible in comparison to the band width observed with the sample in place, the quantity $\Delta\nu$ to be used in Eq. (12) should be obtained from the relation:

$$\Delta\nu = \Delta\nu_{\text{loaded}} - \Delta\nu_{\text{unloaded}}. \tag{14}$$

The term "band width" as used above denotes the difference between the two frequencies for which the displacement amplitude of the system is $\frac{1}{2}\sqrt{2}$ times the value found at resonance, under the condition that the system is excited by a sinusoidal force whose amplitude is independent of frequency.

IV. APPLICATIONS

The method described here may be employed successfully in conjunction with any of the known techniques for measuring the fundamental resonant frequency and the band width of a rod with free ends. A technique that is suitable when the rod is made of magnetostrictive metal has been described in a previous publication.[1] The completely electrostatic technique of Bordoni[3] is also well suited to this application. This method is being used by a group at M.I.T. for the study of deterioration of adhesives.*

In general, this method is suitable for investigating the elastic properties of thin samples whose mechanical impedance per unit area is large compared to the specific acoustic impedance, ρc, of the rod material, provided of course that the samples can be bonded suitably to the rod metal. For the study of samples whose mechanical impedance per unit area is small compared to the specific acoustic impedance of the rod material, it is necessary to place the sample at a position of low impedance (that is, at an antinode of displacement) with respect to the wave pattern in the rod. This case has been discussed in connection with the magnetostriction technique to which reference has already been made.[1]

[2] Slater and Frank, *Mechanics* (McGraw-Hill Book Company, Inc., New York, 1947), p. 32.

[3] P. G. Bordoni, "Metodo Elettroacustico per Ricerche Sperimentali sulla Elasticita," Nuovo Cimento 4, 177 (1947).

* Ordnance Dept. Project No. TB4-601, Professor A. G. H. Dietz, Supervisor.

APPENDIX

More Rigorous Treatment of the Vibration of a Split Rod

It is assumed that the Q of the vibrating system is at least of the order of 10 to 100. The motion, then, is closely symmetrical about a plane normal to the rod and through the center of the viscoelastic sample. This plane of symmetry contains a displacement node which may be replaced by a termination of infinite impedance. The problem resolves into the treatment of a rod whose length is that of one-half of the actual rod, terminated by a rigidly backed viscoelastic sample whose thickness is half that of the actual sample. The analysis is easily performed by the hyperbolic tangent method of Morse[4] and is given below in the terminology established by him.

In the plane joining the sample and the metal rod ($x=0$), the specific acoustic impedance $2(E_2+iE_1)/b\omega$ of the half-thickness viscoelastic sample must equal the transmission-line impedance at this location as given by the function

$$z=\rho c \tanh\pi(\alpha-i\beta_0). \tag{15}$$

Here β_0 represents the value at $x=0$ of the phase parameter β which obeys the relation

$$\beta_{l/2}=\beta_0+l/\lambda. \tag{16}$$

Evidently the dynamic modulus, E_1-iE_2, can be calculated from a knowledge of the quantities α and β_0.

It will be shown that these quantities may be evaluated by measurement of the resonant frequency and the band width of the system.

Let $K\omega$ represent the amplitude of an alternating pressure whose magnitude is proportional to frequency, applied to the free end of the rod at $x=l/2$. The displacement amplitude at $x=l/2$ is $K/z_{l/2}$. Expansion of the hyperbolic tangent Eq. (15) shows that the square of this displacement amplitude is proportional to

$$\frac{A\cos^2\pi\beta_{l/2}+B\sin^2\pi\beta_{l/2}}{B\cos^2\pi\beta_{l/2}+A\sin^2\pi\beta_{l/2}} \tag{17}$$

where $A=\cosh^2\pi\alpha$, $B=\sinh^2\pi\alpha$. It may be verified by differentiation that this squared amplitude is maximum for $\beta_{l/2}=0$. Thus the parameter β_0 is given immediately as

$$\beta_0=-\nu l/c=-\nu/2\nu_0=-\tfrac{1}{2}\left(1+\frac{\delta\nu}{\nu_0}\right), \tag{18}$$

where ν and ν_0 are the frequencies of maximum amplitude for the compound rod and for an unmodified rod of length l respectively.

At the amplitude maximum, the quantity (17) has the value A/B, but this value must reduce to $A/2B$ at either of the two frequencies which define the band width. By setting the expression (17) equal to $A/2B$, we obtain

$$\tan\beta'_{l/2}=\pm(AB/(A^2-2B^2))^{\frac{1}{2}}, \tag{19}$$

where $\tan\beta'_{l/2}$, which equals $\pm l\Delta\nu/2c$, denotes the values of $\beta_{l/2}$ obtaining at the limits of the "pass band." This expression results in the following transcendental equation for α:

$$\tanh^2\pi\alpha/(1-2\tanh^2\pi\alpha)=\tan^2(\pi\Delta\nu/4\nu_0). \tag{20}$$

By means of Eqs. (15), (18), and (20) it is possible to compute the specific acoustic impedance or the complex modulus of the cement layer from measurements of the resonance frequency and the band width of the compound rod.

Since the above equations are unwieldy for purposes of calculation, it is convenient to introduce approximations. Suppose that the fractional frequency change $\delta\nu/\nu_0$ and the fractional band width $\Delta\nu/\nu_0$ are each small compared to unity. In this case the equations reduce to

$$E_1=4\rho l b\nu_0^2(1+\delta\nu/\nu_0)(\delta\nu/\nu_0)/[(\delta\nu/\nu_0)^2+(\Delta\nu/2\nu_0)^2], \tag{21}$$

$$E_2/E_1=\tfrac{1}{2}(\Delta\nu/\delta\nu). \tag{22}$$

These equations are valid for any ratio E_2/E_1. Where this ratio is small compared to unity, as is the case for many cements, Eq. (21) may be replaced by Eq. (7), obtained in the simpler derivation.

The assumption that the driving-force amplitude is proportional to frequency, made in the derivation leading to Eqs. (21) and (22), is introduced solely to achieve a considerable simplification in the derivation and in the results. This condition can be achieved easily in the design of the apparatus, but the distinction involved is unimportant when the resonance is sharp, as will be the case for hard cements. Results identical to Eqs. (21) and (22) are also obtained if it is assumed that the driving-force amplitude is independent of frequency and that the detection device responds to velocity rather than to amplitude.

[4] P. M. Morse, *Vibration and Sound* (McGraw-Hill Book Company, Inc., New York, 1948), second edition.

The Electronic Isograph for Roots of Polynomials

BYRON O. MARSHALL, JR.
Air Force Cambridge Research Laboratories, Cambridge, Massachusetts
(Received August 10, 1949)

An electronic apparatus which assists in the finding of the roots of polynomials is described. The device accepts the coefficients of polynomials of degree 10 or less and yields the absolute magnitude of the roots to a precision of 0.01 relative to the largest coefficient in the polynomial and an angular accuracy of 0.3 degree on the angle of the root.

The method used is the electronic analog of the expression

$$\omega = \Sigma a_n z^n$$
$$= \Sigma a_n r^n \cos n\theta + j \Sigma a_n r^n \sin n\theta.$$

$\sin n\theta$ and $\cos n\theta$ are generated by a motor-driven commutator. The radius r is controlled manually. Hence a plot of ω is placed on the oscilloscope for a fixed r as a function of θ. To find the root r is manipulated until ω crosses the origin in the ω-plane. The value of θ is found by inserting a pip on the intensity of the oscilloscope under a variable phase control. When this pip is moved to the origin of the ω-plane the phase meter indicates the angle of the root in the z plane.

I. INTRODUCTION

THE great interest in the location of the roots of polynomials is partly evidenced by the great number of methods that have been devised for the process. The most common "pencil and paper" methods include the binomial formula for quadratics, the rather awkward formulas for cubics and quartics, Graffe's root squaring method, the various methods of successive approximation (Newton's Birge-Vieta, etc.[1]) and several more complicated procedures. These methods, however useful for low degree polynomials, are laborious and time consuming for greater than fourth or fifth degree. A convenient process for finding the roots of polynomials of perhaps tenth degree or so would have several fold applications: (1) for the solution to the great variety of engineering problems that may be expressed as functions of polynomials, (2) as an aid to large-scale computing machines for which the roots of a polynomials must be isolated before further calculation can be performed, (3) possibly to give new approaches to problems difficult by other mathematical procedures, since any analytic function may be expressed as a power series, many of which converge rapidly, (4) as an adjunct to other mathematical processes: complementary solution to differential equations, the study of transforms, the determination of latent roots of a matrix, and the like.

Several electrical and mechanical methods have been developed for finding the roots of polynomials, some of which are: the Bell Laboratories isograph,[2] the Stibitz electrolytic isograph,[3] and the University of Texas[4] mechanical synthesizer. The system herein described will possess the advantages: (a) of being able to evaluate a polynomial to three significant figures for all values of the complex variables in a matter of minutes, (b) of being extendable to operation with automatic computers, and (c) of allowing additional operations such as Fourier synthesis.

II. THEORY OF OPERATION

Given the polynomial

$$a_n z^n + a_{n-1} z^{n-1} \cdots a_0 = f(z),$$

it is desired to find the values of complex z for which $f(z)=0$. That is, $f(z)$ will have one value for each point $z=x+jy$ of the complex plane. In particular, $f(z)$ will be zero for some values of z (the roots) and the number of roots will equal the degree of the polynomial. By de Moivre's theorem we can set

$$z = re^{j\theta} = r(\cos\theta + j \sin\theta) \quad r, \theta \text{ real}, \ \geq 0$$

and the polynomial becomes

$$\sum_0^n a_k r^k e^{jk\theta} = f(z).$$

Let θ now be varied periodically from 0 to 2π at a frequency $\omega/2\pi$ cycles per second. It can be seen that the kth term may be represented by a sinusoid of frequency $k\omega/2\pi$ amplified or attenuated by the amount $a_k r^k$. It is possible to investigate $f(z)$ for all values of z by summing these weighted sinusoids and varying r from zero to infinity continuously. Also

$$f(z) = \sum_{k=0}^{n} a_k r^k \cos k\omega t + j \sum_{k=0}^{n} a_k r^k \sin k\omega t,$$

and a root will obtain when both

$$\sum_{k=0}^{n} a_k r^k \cos k\omega t = 0$$

$$\sum_{k=0}^{n} a_k r^k \sin k\omega t = 0.$$

[1] J. B. Scarborough, *Numerical Mathematical Analysis* (Johns Hopkins Press, Baltimore, Maryland, 1930).
[2] R. L. Diezold, "The isograph, a mechanical root finder," Bell Lab. Record (1937).
[3] G. R. Stibitz, "An electric root finder," Memo., Burlington, Vermont.
[4] S. Leroy Brown and R. G. Packard, Phys. Rev. 73, 646 (1948).

In other words, both real and imaginary components must be zero simultaneously. This situation and the value of $f(z)$ for all z are indicated by the isograph.

Physically, then, the method of solution consists of (1) generating the required sinusoids, constant in phase relationships and amplitude, (2) multiplying the amplitude of each sinusoid by its corresponding coefficient, a_k, (3) multiplying each of these quantities by the appropriate r^k, r being continuously variable from zero to infinity, (4) presentation of the sum by some convenient means.

There are many existing techniques for performing each of the above operations; those for the isograph were chosen for ease of operation, reliability and accuracy of circuits.

III. PHYSICAL DESIGN

Although it would be possible to introduce complex coefficients through phase shifting of each frequency, the present model operates with real coefficients only, since a method exists (described in a later paragraph) for obtaining a polynomial with real coefficients from one with complex ones. Let this polynomial be divided through by its largest coefficient. All the remaining coefficients are now one or smaller and so may be obtained electrically with potentiometers. Coefficients are inverted in sign by reversing the phase of the corresponding sinusoid.

Again, it will be shown possible to construct an equivalent polynomial

$$\sum_{k=0}^{n} a_k' r'^k e^{jk\omega} = f(z),$$

such that $r=1/r'$. We can investigate $f(z)$ from 0 to 1 and $f'(z')$ in the same range, covering the whole z plane. If R designates r when $|z|\leq 1$, and r' when $|z|\geq 1$, then R will be everywhere less than one and may also be obtained with potential dividers. The value of R^n may be formed by the ganged potentiometers shown in Fig. 1. Each potentiometer is separated from the following one by buffing amplifiers of gain one. One could use a gang of k potentiometers for the kth frequency, $k+1$ for the $(k+1)$th frequency, and so on, but a more convenient method is to produce d.c. voltages proportional to R^k, k between 0 and n, with a gang of n potentiometers. It then remains to control the amplitude of each voltage of frequency $k\omega$ with the d.c. voltage R^k.

As to the number of frequencies to be generated, or the degree of polynomial that may be generated, ten has been chosen for the electronic isograph, since most frequent polynomials are of this degree or less. Also, as will be shown, the roots of polynomials of higher degree may be found if the roots are sufficiently removed from 1. It is necessary, then to produce the first ten harmonics of the base frequency in constant phase relationships and of controlled amplitude. Some of the many methods of this generation are:

1. Mixing and remixing of the base frequency.
2. Synchronized oscillators or trigger circuits.
3. Magnetic pick up from revolving gears.
4. The photoelectric analogy of (3).
5. The electrostatic analogy of (3).
6. Square wave generators using commutators.

The commutators (6) possess the unique property of having a large output that is a linear function of the applied d.c. voltage, since the d.c. voltage is the amplitude of the square wave output. Other types of generators must be followed by a linear modulator for multiplication of their output by the d.c. voltage, R^k. This consideration eliminated 30 vacuum tubes from the first design of the isograph. Operation using method (2) is being considered for another model.

The ten commutators, Fig. 1, being driven by a 3600 r.p.m. synchronous motor, have been found to have optimum operation with from one to ten segments.

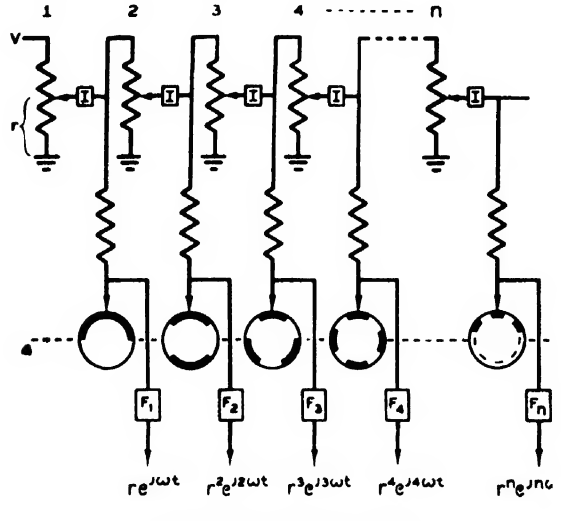


FIG. 1. Generation of terms.

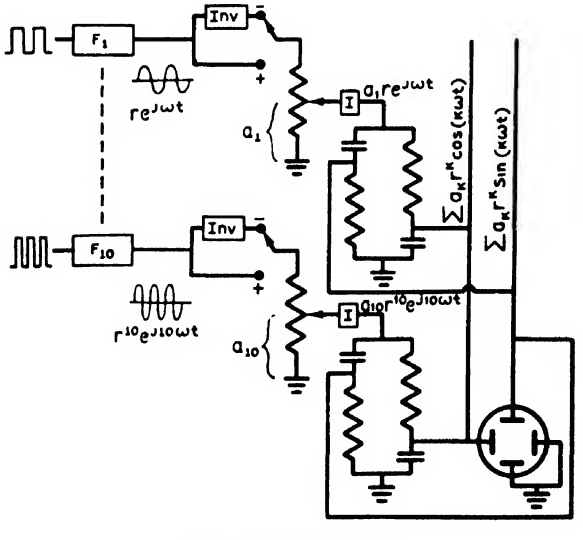


FIG. 2. Block diagram.

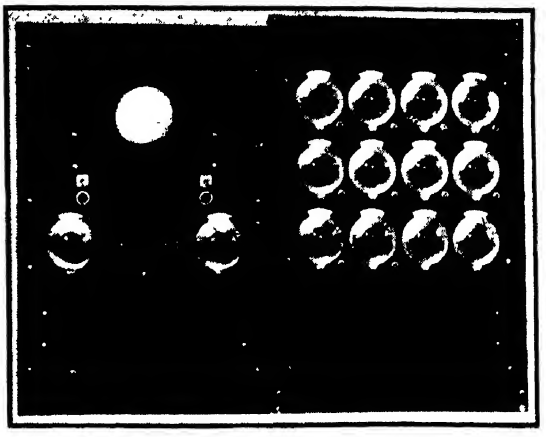
The output frequencies are therefore 60,120 · · · 540 and 600 cycles, each of which will be at least as constant as the line frequency and will have permanent relative phase. The d.c. voltages from the ganged potentiometer are applied to the commutators and each output square wave, proportional to R^k, is filtered to obtain a sinusoidal voltage with harmonics less than 0.1 percent. The potentiometer for multiplication by the fractional a_n may be introduced at a convenient point in the subsequent circuit.

In view of the range of frequencies employed, the type of presentation that immediately suggests itself is an oscilloscope. It is convenient to separate the polynomial into:

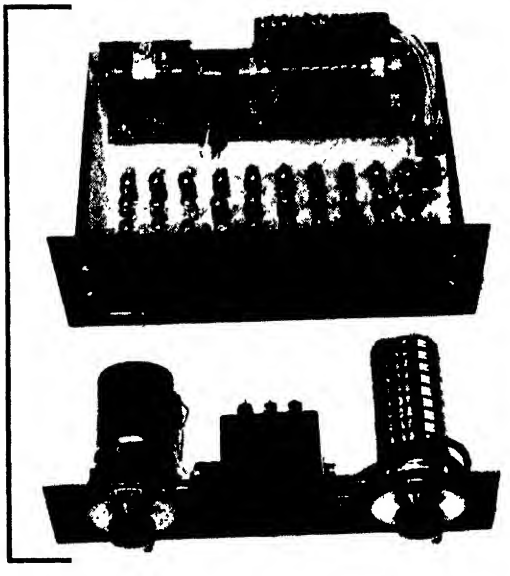
$$f(z)=\sum^{n} a_k R^k e^{jk\omega t}=\sum^{n} a_k R^k \cos k\omega t+j\sum^{n} a_k R^k \sin k\omega t,$$

which may be achieved electrically by splitting each frequency into $\sin k\omega t$ and $\cos k\omega t$ and summing the ten of each in linear mixers. If the sum of the cosine voltages is supplied to the horizontal deflection circuit of the oscilloscope and the sines to the vertical, it is seen that a root of the polynomial will occur when the trace of $f(z)$ crosses the zero point of the oscilloscope. Such a representation has the convenient properties that: (a) The complex value of $f(z)$ is the vector from the origin to that particular point of the trace, (b) the region of greatest interest occurs at the center of the scope face, and hence the presentation may be greatly amplified for greater precision, and (c) since the parameter of the presentation is θ, varying periodically, a particular θ may be indicated by a marker in the proper phase relationship to the fundamental frequency. In the isograph, this marker is produced with a differential selsyn to shift the phase of the fundamental by any given angle, a clipper and peaker and then intensity modulation of the trace with the peaked pulse. In addition, the vector from the origin of the oscilloscope face to a point of the trace is the value of $f(z)$ at that setting of R and the indicated θ.

Using the elements described above, the electronic isograph appears as shown in the block diagram, Fig. 2. It can be seen that the channel for each frequency consists of rather elementary circuits, linear throughout. The zero degree term, the constant in the polynomial, is introduced by horizontal and vertical displacements of the oscilloscope pattern relative to the etched axes. Since all potentiometers have errors of less than 0.07 percent and all circuits are linear, it is possible to align the channels to evaluate a tenth degree polynomial with a constant term to three significant digits. Figure 3(a) shows the physical appearance of the isograph.



(a)



(b)

FIG. 3. Appearance of isograph

IV. OPERATION

Basic Function—Zeros of Tenth Degree Polynomial With Real Coefficients

The only operation that must be performed externally is to divide the polynomial by its largest coefficient, the new polynomial having the same roots as the original. These coefficients are now entered on the vernier coefficient dials, which can be set with an accuracy of one part in two thousand. The values of $f(z)$ for $|z|$ from zero to one are investigated by the R potentiometer through this range. The oscilloscope presentation gives at once a trace representing the values complex $f(z)$ for all z lying on the circle $|z|=R$ for the particular R entered on that dial. A root will be indicated by the trace crossing the origin of the presentation, the magnitude of the root being the R setting when this situation obtains. The argument, or phase angle, of the root may be read from the θ-scale after the θ-marker has been placed on the portion of the trace at the origin. The precision of the root magnitude measurement is 0.05 percent and of the argument, 0.1 degree, which, of course, is one limit of the attainable accuracy.

To investigate the polynomial for $|z|$ greater than one, a slightly different procedure is employed. If the transformation

$$z = 1/Z = 1/(re^{j\theta})$$

is made, the polynomial becomes

$$\sum_{k=0}^{n} a_k z^k = \sum_{k=0}^{n} a_k Z^k = \sum_{k=0}^{n} a_k (r^k e^{jk\theta})$$

which may be multiplied by $r^n e^{jn\theta}$ to give

$$\sum_{k=0}^{n} a_k r^{n-k} e^{j(n-k)\theta} = a_n + a_{n-1} r e^{j\theta} + \cdots + a_0 r^n e^{jn\theta}.$$

This transformation is performed simply by inverting the order of the coefficients, allowing the completion of the investigation of the z plane from 1 to ∞. The reciprocal of r, r', and the argument are again read from the R and θ scales, respectively.

Since the figure traversed in the z plane is a circle with a radius r (determined by the ganged potentiometer setting), the figure on the oscilloscope face (the w plane) is the transform of the circle. From complex variable theory, a line through a branch point of a function is transformed into a cusp. The branch points are the roots of the derivative, and another method of finding roots presents itself. It is merely necessary to integrate the polynomial (a simple multiplication by integers) and observe the occurrence of cusps on the w plane, as shown by the seven simultaneous roots of Fig. 4.

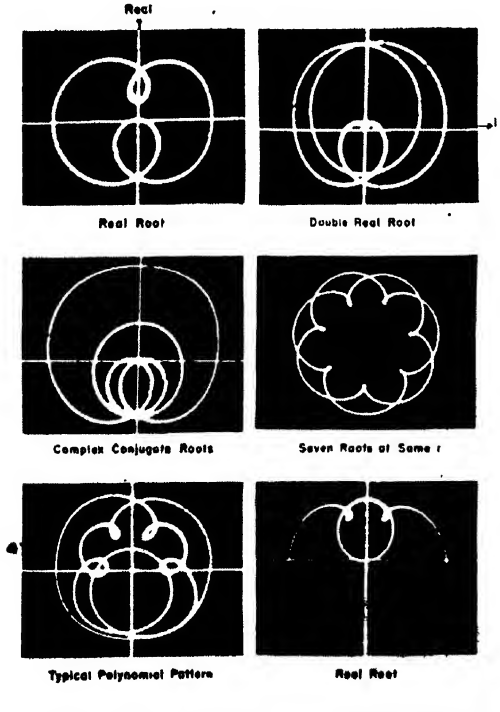


FIG. 4 Real, complex conjugate, double roots, etc.

The accuracy of the isograph is greater near $(z)=1$, so that large or small roots should be brought to this neighborhood by constructing a polynomial in z', z' being defined as Mz when M is some large (or small) real number.

Many of the properties of the polynomial may be observed on the $f(z)$ presentation. The investigation of $f(z)$ is carried out over circles about the origin in the z plane, since r is constant at any instant and θ is being scanned. Any polynomial, $P_n(z)$, may be expressed as the product

$$P_n(z) = A(z-p_1)(z-p_2)\cdots(z-p_n),$$

where $p_1 \cdots p_n$ are the roots of $P_n(z)$. The quantity $(z-p_k)$ will change its argument by 2π as z moves around a circle enclosing p_k. If the circle C encloses m roots, the argument of $P_n(z)$ will change by $2m\pi$ for each circuit or z on C, or the trace of $P_n(z)$ will enclose the origin m times. In other words, the number of loops made by the trace of $P_n(z)$ about the origin of the oscilloscope is equal to the number of roots of absolute value less than the r then entered in the isograph.

The character of each root may also be determined from the presentation. Real, complex conjugate, double roots, etc., have the characteristic appearance shown in Fig. 4. It is also possible to determine some of the properties of the derivative at the root from the presentation.

If it is desired to obtain greater accuracy, the polynomial may be differentiated and the resulting $f'(z)$ entered in the isograph as a new polynomial to be evaluated at the R and θ of the root's approximation given by the first process. One can then apply Newton's formula:

$$z_1 = z_0 - [f(z_0)/f'(z_0)],$$

where z_0 is the first approximation, z_1 a better value. It is necessary to calculate the value of $f(z_0)$ by some auxiliary means; but since each application of Newton's formula usually doubles the number of significant figures, the additional calculation is efficient.

Roots of Polynomials with Complex Coefficients

A necessary and sufficient condition for a polynomial to have real coefficients is that its roots be distributed symmetrically about the real axis in the z plane. A polynomial with complex coefficients may be changed to one with real coefficients by introducing the conjugates of the roots of the original. Since it may be shown that

$$f(\bar{z}) = \langle f(z) \rangle_{Av}$$

the conjugate roots may be introduced by forming the product

$$F(z) = f(z) \cdot \langle f(z) \rangle_{Av}.$$

The new polynomial will in general be of twice the degree of the original and have twice the number of roots. The additional roots are the conjugates of the

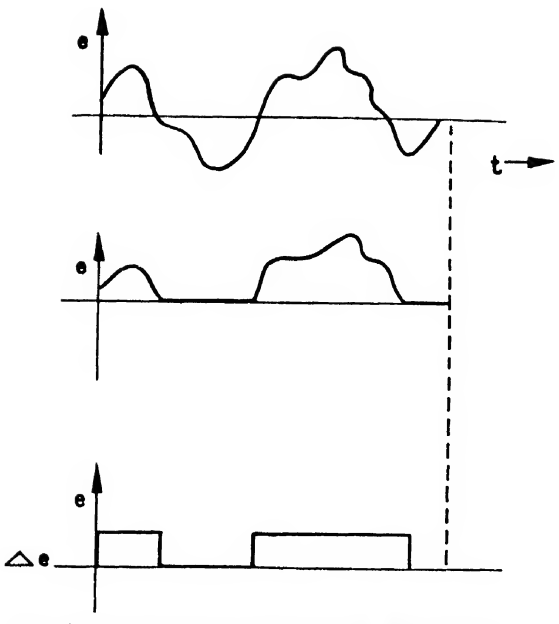


FIG. 5.

desired ones and may be eliminated by substitutions in the original polynomial. The roots of $F(z)$ are found by as described in the first paragraphs of Section IV.

Polynomials of Degree Higher Than Ten

A polynomial of degree $10+n$ may have all of its roots found by the isograph if n of its roots are sufficiently removed from one. The upper ten terms will give perhaps ten of the roots correct to three places if the roots are greater than 1.6 or less than 0.6, assuming reasonable coefficients. The terms neglected will contribute little to the root determination since $(1.6)^{10}$ is greater than 1000 and $(0.6)^{10}$ is less than 1/1000, meaning that the lower terms' contributions are outside the accuracy of the instrument. The larger roots, discovered by using the upper ten terms, may be divided out of the original polynomial, resulting in one of lower degree which may then be solved by the isograph for the remaining roots.

Again, if the roots are too close to one for the above conditions to obtain, it is possible to remove them from this neighborhood by defining a new variable or by using Horner's method

V. EXTENSIONS

It is possible to give a presentation much easier to interpret by mapping the polynomial over the z plane. This is a familiar type of representation and would be much more recognizable by mathematicians and engineers. It is difficult to accomplish but a method which would indicate the important points of the map would be valuable. One way of doing this, which it is believed will give considerable information to the engineer, would be as follows.

The previously mentioned oscilloscope presentation was developed by having a voltage on the horizontal plates proportional to the sum of the cosines of the various terms of the polynomial and was some periodic voltage centered about 0. The voltage on the vertical plates is the same type representing the sum of the sines. If each voltage were clipped at zero and at some small voltage above zero (see Fig. 5), and the resulting rectangular wave form were placed on the z axis of an oscilloscope while at the same time tracing on the face of the oscilloscope the circle corresponding to z magnitude$=r$ (r the setting of the r potentiometer), then this intensity modulation defines the regions of positive real (or imaginary) $P_n(z)$. This means that as r is varied from 0 to 1 there is obtained on the face of the oscilloscope a set of zones corresponding roughly to the surface of real (or imaginary) $P_n(z)$. The edges of these zones, of course, represent the lines real $P_n(z)=0$ (or imaginary $P_n(z)=0$). Although this is an approximation to the surface $P_n(z)$ mapped above the z plane, the significant points, roots, and branch points are given accurately. It is possible to put both surfaces on the face of the oscilloscope by simply adding the two intensity modulating voltages. There are then displayed on the z plane, intersections of the zero lines indicating the roots, and other shapes, the branch points of the

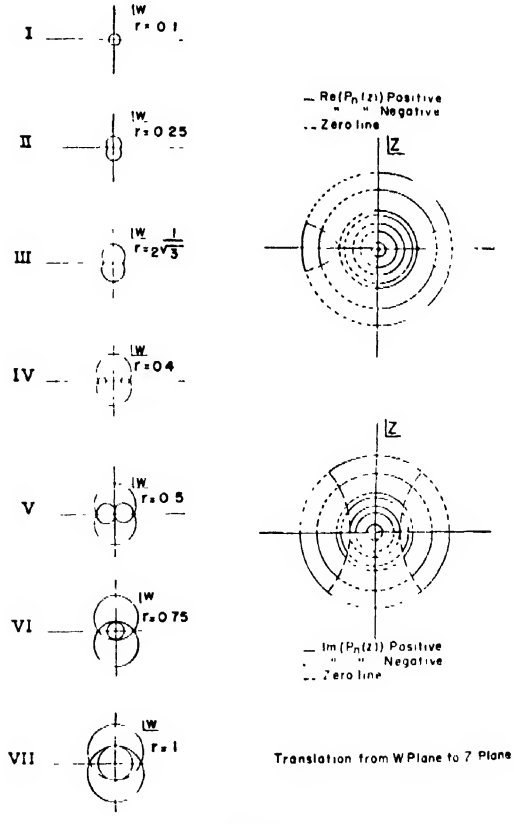


FIG. 6

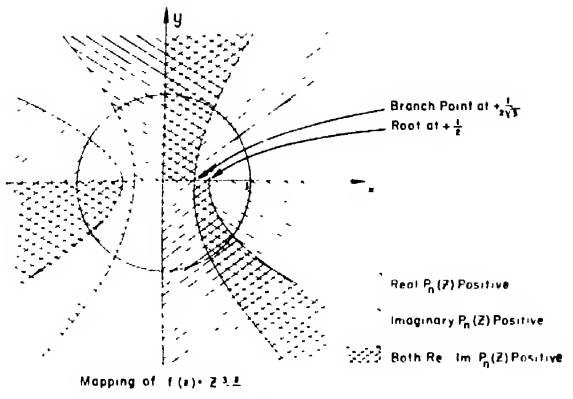


FIG. 7

polynomial, and it will be possible to trace the behavior of these significant points as coefficients of the input polynomial are varied. Figure 6 is an illustration of what would be obtained using this presentation on the polynomial $z^3-z/4$. On the left are the series of illustrations representing the transforms of circles in the z plane to the w plane which are the usual isograph presentation. To the right are the two presentations representing the surface real $P_n(z)$ and imaginary $P_n(z)$ and Fig. 7 is the sum of the two upon which may be seen the roots and the branch points of this function. It is believed that this is the first attempt at mapping a function electronically. Instrumentation of this presentation is presently being completed.

Also, the coincidences of zeros in both real and imaginary parts of $P_n(z)$ may be detected electronically and presented on the z axis of the oscilloscope representing the z plane. The horizontal and vertical voltages again sweep the oscilloscope beam in a circle of $r=z$. The trace is brightened at the roots and the presentation is one in which the roots are indicated by bright spots on the z plane.[5] Again it will be possible to trace the movement of the roots as coefficients are varied.

VI. CONCLUSION

The electronic isograph can give rapid solutions to reasonably well-behaved polynomials to three places. With the addition of a small amount of paper work it is possible to find the zeros of all but a very small number of polynomials. It is hoped that the isograph will prove extremely valuable in servomechanism and other fields where the problem may be represented as a polynomial and that theoretical analysis of engineering problems will be furthered by its use.

VII. ACKNOWLEDGMENTS

The author wishes to thank co-workers of Special Studies Laboratory, A.F.C.R.L., for suggestions and assistance in the construction of this first model, especially Mr. Eugene F. Grant, whose suggestions inspired the project.

[5] Presentation of this type used by O. Vierlins, Germany. No reference available.

The Cadmium Sulfide X-Ray Detector*

RUDOLF FRERICHS
Department of Physics, Northwestern University, Evanston, Illinois
(Received September 12, 1949)

The CdS x ray detector is compared with the photo-multiplier x-ray detector. Reduced to the same solid angle the sensitivity of the CdS x-ray detector is of the same order of magnitude as the sensitivity of the photo-multiplier x-ray detector. As the internal amplification of the primarily released electron currents in CdS is of the order 10^3 to 10^4 compared with the amplification in the multiplier tube of $\frac{1}{2}\times10^6$ the conversion process of x-ray quanta in slow electrons in CdS must be more efficient than the corresponding processes in the combination crystal phosphor screen photo-multipher. Calibration curves show a linear relation between photo-current and intensity of x-rays at low irradiation levels. At high irradiation saturation effects occur. Oscilloscope curves of the response at 60 x-ray pulses per second are given. At strong irradiation the photo-current over a long time increases in a completely reversible way. X-ray intensities as low as 150 quanta CuK_α per second are registered with a circuit employing a WL 759 trigger tube.

INTRODUCTION

THE measurement of x-rays with a combination of a crystal phosphor screen and a photoelectric cell has been introduced by Herz in 1936.[1] Since this time every progress in the development of new phosphor screens or photo-cells has increased the sensitivity of this method. A great step was achieved in 1941, when Hodges and Morgan[2] introduced the photo-multiplier tube x-ray detector. Subsequently this combination

* This investigation was sponsored by the U. S. Navy, Bureau of Ships.

[1] R. H. Herz, Fortschr. Gebiete Röntgenstrahlen 54, 6 (1936). It is interesting to note that in his first papers Röntgen tried to determine the distribution in space of the x-rays around the tube by photometrie of barium platino cyanide screens. Now the expensive and unstable barium platino cyanide screens have generally been replaced by crystal phosphors.

[2] R. H. Morgan, Am. J. Roent. Rad. Ther. 47, 777 (1942).

has been thoroughly investigated[3] and the photo-multiplier x-ray detector is now a sensitive and reliable x-ray measuring device.

On the other hand, it was shown in 1946 by the author that synthetic CdS crystals show a large photo-conductivity under irradiation with x-rays.[4] In the photo-multiplier x-ray detector the energy of the absorbed x-rays is transformed at first in energy of slow electrons of a few ev. These electrons raised into the conduction band, subsequently recombine with the activating impurities and produce photons. A part of these photons is absorbed in the photo-cathode of the multiplier and releases a small yield of photo-electrons. These in turn release increasing cascades of secondary electrons at the consecutive stages of the multiplier. One photo-electron, liberated from the cathode, yields up to $\frac{1}{2}\times10^6$ electrons in the anode circuit. In the CdS crystal the same primary process of raising a number of slow electrons into the conduction band takes place. But in contradistinction to the man-made photo-multiplier arrangement the nature of the crystal provides the subsequent amplification of these electron currents.[5]

While in the photo-multiplier the rate of amplification in the different stages can be easily investigated, in CdS the amplification factor is only approximately known. Estimates made under rather plausible assumptions yield values of 10^3 to 10^4. Exact figures could be obtained by a simultaneous study of the light emission and the photo-conductivity of the irradiated CdS crystal. The light emission is determined by the number of electrons in the conduction band which recombine with the activating impurities. The photo-conductivity is related to the lifetime of the positive stationary space charges at the activating impurities. They make the presence of an additional electron in the conduction band possible.[6] As their lifetime is large, compared with the time one electron needs for passing through the crystal, a chain of up to 10^4 electrons cross the crystal, until the positive charge is neutralized by recombination. This amplification of the primary photo-current appears to be less favorable in the CdS crystal than in the multiplier tube. The primary processes, however, are simpler and are apparently connected with smaller losses than the corresponding processes in the combination phosphor screen photo-multiplier.

The purpose of this investigation to compare the two devices under identical conditions in order to find out for which purpose the one or the other might be best applied.

TABLE I. X-ray absorption in CdS crystals.

Wave length	Crystal thickness: 99 percent absorption	Crystal thickness: 99.9 percent absorption
0.13A	7.4 mm	11.1 mm
0.20	2.16	3.24
0.31	0.66	0.99
0.40	0.36	0.54
0.45	0.25	0.37
0.47 (K edge)	1.30	1.95
0.51	1.06	1.59
0.61	0.62	0.93
0.71	0.42	0.63
0.83	0.28	0.42
0.92	0.20	0.30
1.04	0.14	0.21
2.09	0.02	0.03
3.02	0.009	0.013
3.72 (L edge)	0.028	0.042
6.14	0.008	0.012
9.86	0.002	0.003

D.C. MEASUREMENTS OF THE PHOTO-CURRENT

Measurements were performed with copper target tubes provided with beryllium windows.[7] The tubes were used with half-wave and with full wave rectification but without condensors. They furnished 60 (half-wave), 120 (full wave) x-ray pulses per second.

For a comparison at higher intensities the crystal, as well as the photo-multiplier x-ray detector, were placed in the beam at a distance of 50 cm from the target. For this informatory experiment no filtering other than 1.5-mm beryllium window was used. The photo-multiplier tube was used with 102 volts across each dynode corresponding to an over-all amplification of the electron current by a factor of 1.3×10^6. The d.c. currents were measured with a sensitive micro-amperemeter (full scale 5, 50, 500, 5000×10^{-6} amp.). The photo-multiplier tube was shielded in a brass container, and a fluoroscopic screen, Patterson Type B, was attached at the tube surface behind a black, paper window. The dark current of the tube was about 1×10^{-7} amp. When the multiplier was irradiated without the phosphor screen at 24 kv and 10 ma, the current rose to $3.8\pm0.2\times10^{-6}$ amp. The crystals to be investigated have sensitive areas of the order of 1 mm^2. Accordingly, a comparable size of the phosphor screen was used. The phosphor layer was scratched from the supporting cardboard leaving an area of 1×5 mm. Under these conditions the current obtained under x-ray irradiation amounted to 12.5×10^{-6} amp., but quickly dropped to 9×10^{-6} amp. because of the fatigue effects extensively discussed by Marshall, Coltman, and Hunter.[3]

The CdS crystal showed a bright red luminescence under irradiation with filtered ultraviolet. It is known

[3] H. M. Smith, Gen. Elec. Rev. **48**, 13–17 (1945). Marshall, Coltman, and Hunter, Rev. Sci. Inst. **18**, 504 (1947). J. W. Coltman, Proc. I. R. E. **37**, 671 (1949).

[4] R. Frerichs, Naturwiss. **33**, 281 (1946). R. Frerichs, Phys. Rev. **72**, 594 (1947).

[5] Natural amplification phenomena will play a more important role as the number of electronic devices applied in complicated mechanism (electronic calculators, electronic brains) reaches values which make it necessary to reduce the size and the labor involved in the individual device. An example is the recently developed transistor, which can replace complicated electron tubes. See furthermore: V. K. Zworykin and E. G. Ramberg, *Photoelectricity*, p. 469.

[6] R. Frerichs, Phys. Rev. **76**, 1869 (1949).

[7] The author is greatly indebted to Dr. J. C. M. Brentano, who made the equipment of his x-ray laboratory available to him.

TABLE II. Current through crystal and multiplier as a function of the x-ray intensity.

X-ray tube Potential	X-ray tube Current	r per min.	Current $\times 10^{-6}$ amp. Crystal	Current $\times 10^{-6}$ amp. Photo-multiplier
70 kv	2.1 ma	0.063	0.487	0.463
65	2.1		0.450	0.390
60	2.1		0.425	0.325
50	2.1	0.031	0.375	0.270
45	2.1		0.306	0.220
40	2.1		0.250	0.165
35	2.1		0.175	0.110
30	2.1	0.014	0.100	0.068
25	2.1		0.037	0.037
20	2.1		0.005	0.011
15	2.1		0.000	0.002

(0.5 mm Cu filter)

that luminescent CdS crystals yield large amplification.[8] The sensitive area between the Aquadag electrodes was 0.3×1.5 mm^2. It was mounted in a brass tube with a black, paper window. The leads were well insulated and shielded. At an applied potential of 280 volts the dark current remained smaller than 1×10^{-8}; it could not be observed with the microamperemeter used. Within five seconds the current rose under irradiation to $27 \pm 1 \times 10^{-6}$ amp. The amount of energy absorbed in the phosphor layer or in the crystal was not determined. The thickness of the crystals was not uniform. These crystals show a ribbon-like structure. Therefore, exact measurements of the absorption are difficult to perform. It was observed, however, with a fluoroscopic screen that only a small amount of incident soft x-rays passed through the crystal at its average thickness of 0.2 mm. Table I shows the thickness which is necessary to absorb 99 or 99.9 percent of the incident irradiation at different wave-lengths. The values are calculated from the mass absorption coefficients of cadmium determined by Laubert and Biermann.[8a] In the range between 1 and 2A no data were available.

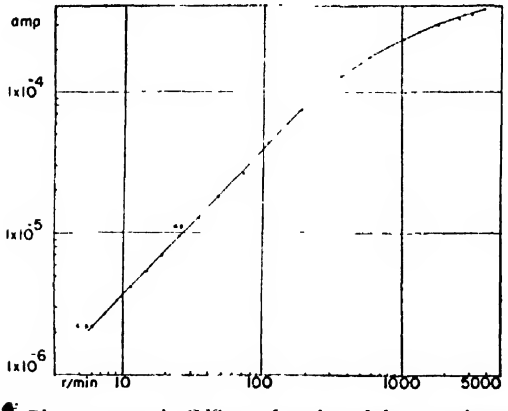


FIG. 1. Photo-current in CdS as a function of the x-ray intensity.

[8] This is to be expected from the theory outlined in reference 6. High luminescence proves that sufficient activating impurities are present which furnish the positive stationary space charges necessary for the compensation of the transported negative charges.

[8a] S. Laubert, Ann. d. Physik **40**, 553–578 (1941). H. H. Biermann, Ann. d. Physik **26**, 740 (1936).

It was not attempted to consider the influence of the scattering coefficient which is contained in these measurements but which is usually much smaller than the coefficient of true absorption. In contradistinction to other photo-conductors with larger dark currents, it is not necessary to chose the thinnest crystals. For short wave-lengths it is recommended to irradiate the ribbon-like crystals ($a > b > c$) through the surfaces $a \times c$ or even $b \times c$ in order to obtain high absorption. Thus, the effective area can be small, which is an advantage for measuring narrow bundles of x-rays.

From these experiments it follows that reduced to the same solid angle which corresponds to an effective area of 0.5 mm^2, the photo-currents obtained in CdS are one order of magnitude larger than those obtained with the multiplier.

The luminescent screen used (Patterson B) is not the most efficient. The photo-multiplier responds about three times better to the blue luminescent screen, Type C. It should be possible to use the entire light emitted by the phosphor. For instance, with a spherical mirror such as Coltman[3] used. Furthermore, the sensitive area of the phosphor screen generally should be larger (50 mm^2) in order to increase the solid angle without increasing the dark current.

On the other hand, the potential applied at the CdS crystal can be increased until the dark current reaches values comparable with the photo-multiplier dark current. Furthermore, it is possible to assemble different crystals in one plane and to connect the electrodes parallel.

In conclusion we might state that using d.c. currents the CdS crystal can be made one order of magnitude more sensitive per unit area than the photo-multiplier x-ray detector. The experiments show that the mechanism of transforming x-ray quanta into slow electrons in CdS is apparently much more effective than the transforming of x-rays→slow electrons→photons→photo-electrons at the photo-cathode since a deficiency in amplification of 10^{-2} to 10^{-3} is more than compensated.

THE PHOTO-CURRENT IN CdS AS A FUNCTION OF THE X-RAY IRRADIATION

A few measurements were made to study the photo-current in CdS as a function of the irradiation. It was previously observed by the author[6] that at high irradiation levels—for instance, with gamma-rays—the current produced in CdS crystals is proportional to the square root of the intensity of irradiation. This relation follows theoretically from the recombination processes involved in these secondary photo-currents. In equilibrium after prolonged and strong irradiation, the number n of the electrons in the conduction band and thus the current is determined by

$$dn/dt = aI - bn^2 = 0.$$

I is intensity of the impinging radiation; a is a factor which contains the absorption in the crystal and the

efficiency of conversing x-rays in slow electrons. The factor b determines the recombination rate between electrons in the conduction band and activating impurities in the lattice. At the first approximation trapping and releasing of electrons, which play such an important role in persistent phosphorescence, do not need to be considered in the equilibrium state. The number of the electrons which are trapped per second is equal to the number of electrons which are released from the traps into the conduction band per second. We make the assumption that first-order processes do not take part. Nail, Pearlman, and Urbach[9] recently pointed out that at low irradiation levels in crystal phosphors first-order processes such as energy transfer from trapped electrons to the lattice by direct transition can occur. Under certain circumstances a region of proportionality to the square root of the exciting intensity is situated between linear regions at low and extremely high intensities. Therefore, the dependence of the photo-current from the incident radiation should not be linear at high intensities. It might be pointed out here that the same is to be expected for the photo-multiplier x-ray detector if a crystal phosphor is used.[10] The combination with a fluorescent emitter, for instance, barium platino cyanide, yields a linear characteristic as fluorescent substances do not show any saturation effects.[11]

Some measurements were taken with the copper target tube at 40 kv and 20 ma. The crystal was used with 22.5 volts. The distance between crystal and target was varied from 8 to 147 cm. The results are plotted in the log-log diagram Fig. 1. The abscissa contains the true x-ray intensities. According to the manufacturer of the apparatus, at 40 kv and 20 ma the intensity of the unfiltered x-rays at 50-cm distance from the target amounts to 100 r per min. According to the measurements of Day and Taylor,[12] considerable corrections have to be applied at the simple distance square relation, if the tube potential is low and if no filters are interposed. The intensity values in Fig. 2 are corrected, using Fig. 4 of Day and Taylor, assuming a value of 40 kv and 1.5-mm beryllium window. The curve covers the range 10 r per min. up to 5000 r per min. It is straight at low values and becomes curved above 100 r per min. The tube used in our experiments with 40-kv peak current emits a spectrum, the distribution of which shifts periodically with the alternating voltage applied. In Fig. 1 the size of corrections is marked under the conservative assumption that the actual average voltage would be as low as 20 kv. Another possible correction, the slow change of the photo current with time, will be discussed later.

TABLE III. Current through crystal and multiplier at different filtering.

X ray tube Potential	X ray tube Current	Thickness of Al filter	Current ×10 ^{8} amp. Crystal	Current ×10 ^{8} amp. Photo multiplier
60 kv	2.2 ma	0	0.475	0.355
60	2.2	$\frac{1}{4}$ in	0.200	0.080
60	2.2	$\frac{1}{2}$ in	0.100	0.032
60	2.2	$\frac{3}{4}$ in	0.042	0.012
60	2.2	1 in	0.017	0.005
60	2.2	5/4 in.	0.005	0.002

Measurements at low intensity levels have been performed by Mr. J. E. Jacobs of the Department of Electrical Engineering, Northwestern Technological Institute.[13] The measurements were taken with a tungsten target tube. The x-rays in Table II are filtered by 0.5 mm Cu. The r values were determined by a Victoreen r meter. The crystal had a sensitive area of 3×1.5 mm^2; the field strength applied was low and equal 90 volts per 3 mm. The currents were measured with a d.c. amplifier. The photo-multiplier tube 931A was selected from a large number. The potential at the single dynodes was 105 volts. A Patterson E_2 industrial screen was used behind a lead aperture of 3×1.5 mm. This limited the solid angle to the value of the crystal.

With increasing potential the efficiency becomes smaller similarly as the efficiency of an ionization chamber decreases with increasing hardness or radiation. As the radiation was not homogeneous, it is not possible to construct a curve, sensitivity *versus* wave-length.[14] The general shape of such a curve should indicate that ionization in air and photo-conductivity in CdS are not parallel, if the hardness of the x-rays is changed.[15] Furthermore, the K absorption edge of cadmium at 0.460A should produce a discontinuity in this curve analogous to the discontinuity produced in the photographic plate due to the silver K edge. Similar

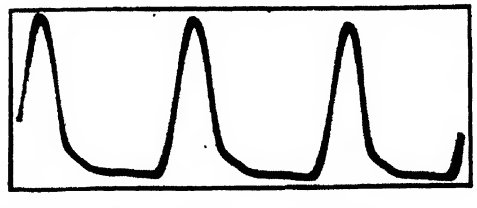


FIG. 2. 30 kv, 20 ma, distance target-crystal=17 cm, CdS crystal area=0.3×1.5 mm, applied potential=1.55 volts, current amplitude=5.5×10^{-8} amp.

[9] Nail, Perlman, and Urbach, "Preparation and characteristics of solid luminescent materials," Cornell University, Symposium (1946), p. 190.

[10] O. Glasser, *Medical Physics* (Chicago 1944), p. 1293. Figure 1, Patterson Type B screens show an increase of brightness which is less than proportional to the impinging radiation at high intensities (3–4 r per min.) especially at lower x-ray tube potentials.

[11] W. de Groot, Physica 6, 393 (1939).

[12] F. H. Day and L. S. Taylor, J. Research Nat. Bur. Stand. 40, 393 (1948).

[13] The author is much indebted to Mr. Jacobs for the facility of using these values in the present investigation.

[14] It might be useful to determine this dependency, for instance, with the powerful filter method developed by H. Kustner, Zeits. f. Physik 70, 324 (1931). According to his measurements monochromatic radiations $\Delta\lambda=0.005$A with intensities from 5×10^{-4} to 27×10^{-3} r per min. can be produced in the wave-length region 0.128 to 2.50A.

[15] For a comparison of the effects produced by x-rays of different wave-lengths at ionization in air, at fluorescent screens and at the photographic plate see: G. L. Clark, *Applied X-Rays* (New York, 1940), p. 68, Fig. 48.

discontinuities are to be expected with the crystal phosphor photo-multiplier x-ray detector at the zinc and cadmium K absorption edges.

Table III contains further measurements with the photo-multiplier and the CdS crystal obtained by Mr. Jacobs. The potential was constant, and different thicknesses of aluminum filtering were applied. Crystal and photo-multiplier were used in 150-cm distance from the tungsten target tube.

THE SPEED OF RESPONSE

As the photo-multiplier can follow light pulses of the order of 10^{-8} sec. the speed of the photo-multiplier x-ray detector is determined by the properties of the screen used. Fast responding phosphors (calcium tungstate) show, according to Marshall, Coltman, and Hunter, a fast and exponential decline of the photo-current, after irradiation with x-ray pulses of 1 μsec. duration. The somewhat persistent sulfide screens, Patterson B and C, are slower. According to their oscillograms the response of the phosphor screen does not match the short x-ray pulse applied. It is not possible to determine its exact shape from such measurements.

Identical experiments could not be performed with the CdS crystal, as the necessary radar pulsing equipment was not available. However, the usual chopping of the x-ray output with 60, respectively, 120 pulses per second, which is obtained by one way or two way rectifiers is very useful for the oscilloscopic observation of the currents in CdS.

A comparison was made of the current *versus* time curves obtained with the photo-multiplier x-ray detector and the CdS crystal at the half-wave tube. The comparison of Figs. 2 and 3 shows that the reproduction of the x-ray intensity curve obtained with these two devices is nearly identical. The declining branch in the photo-multiplier curve approaches the ideal curve more perfectly,[16] but also at these low frequencies the slow decline of the phosphor in the photo-multiplier can be observed. According to Marshall, Coltman, and Hunter, the screens, Type B and C, are somewhat persistent. This dissymmetry is to be attributed to the afterglow of the phosphor. The curves show that at these rather low frequencies the CdS crystal follows the x-ray intensity distribution with good approximation. Figure 4 has been taken with the CdS crystal connected directly to the input resistance of the oscilloscope without battery. In the polar crystal the x-rays produce a small e.m.f., which can be observed with the oscilloscope, if the irradiation is strong enough.[16a]

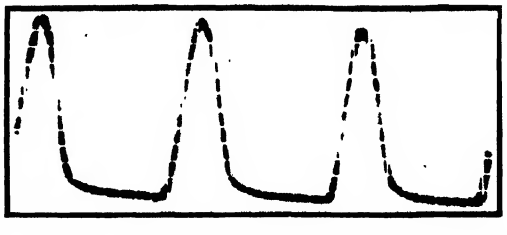


FIG. 3. 30 kv, 8 ma, distance 50 cm, photo-multiplier x-ray detector, screen size = 1 mm², Patterson Type B, dynode potential = 123 volts, current amplitude = 1.3×10^{-7} amp.

[16] The intensity of the afterglow of the phosphor is proportional to the number of recombination processes bn^2. The photo-current is proportional to the number n of electrons in the conduction band. Hence the optical decay is faster than the decay of the photo-current. Verified by L. Gildart and A. W. Ewald in a paper given at the Oak Ridge meeting of the Physical Society (March, 1950).

Curve 4 shows the current in the crystal under weak irradiation. As it has been observed with the luminescence of crystal phosphors, the slope of the raising and declining branch of these curves depends upon the intensity of irradiation. At low intensities the current rises and declines slowly, and the x-ray curve is badly reproduced. One should expect that at low intensities the curve of the photo-multiplier x-ray detector should show some similar shape.

From these experiments it can be concluded that for x-ray pulses of 60 per sec. the CdS crystal can replace the photo-multiplier x-ray detector.

FIG. 4. 40 kv, 20 ma, 17 cm distance, CdS crystal without potential applied, crystal photo-current amplitude = 4.8×10^{-9} amp.

EFFECTS DUE TO PROLONGED IRRADIATION

The fatigue effects observed in photo-multiplier x-ray detectors are due to the multiplier tube. With sufficiently low levels of irradiation they can be greatly reduced. The phosphor screen used with the photo-multiplier shows no change in its sensitivity with prolonged irradiation. An experiment was made to find out whether small changes in the brightness of the phosphor screen take place at prolonged irradiation with intense x-rays. The detector was irradiated with 36,000 r at 60 pulses per sec. In order to avoid fatigue of the multiplier the potential at the dynodes was interrupted during this irradiation. The currents observed at the beginning and at the end of this irradiation agreed within 0.6 percent. This proves that the Patterson B screen does not undergo any observable changes in its luminosity. Such effects could be due to a slow rearrangement of the electrons in the traps of different depths.

A CdS crystal was adjusted so that it showed a photo-current of 85×10^{-6} amp., after a few seconds of irradiation. This current slowly increased during 11 hr. of continuous irradiation until it finally reached 182×10^{-6} amp. The irradiation during this period amounted to 2.5×10^{5} r. The x-rays were pulsed (60 per sec.) and the potential at the crystal was kept constant. The dark current remained throughout this experiment below 10^{-8} amp. After about 12 hr. recovery the crystal

[16a] T. Broser and R. Warminsky, Zeits. f. Naturforschung 5A, 63 (1950).

FIG. 5. 40 kv, 20 ma, 130 cm distance, CdS crystal, potential applied at crystal = 45 volts, current amplitude = 1.2×10^{-8} amp.

reassumed its original value, corresponding to 85×10^{-6} amp. at that particular level of irradiation; and no permanent changes remained. This slow change of the resistance has been investigated with gamma-rays.[6] It is attributed to a rearrangement in the occupancy of deep and shallow traps during prolonged and intermittent irradiation.

The slow increase in the final current has to be considered at all d.c. measurements with these crystals. In the measurements of Fig. 1 the currents were always read after 5 sec. irradiation.

It has been shown by the author that the original high dark resistance can be quickly restored by irradiation with infra-red light.

MEASURING OF SMALL X-RAY INTENSITIES WITH CdS CRYSTALS

Measurements of very weak x-rays were made using a trigger tube (Westinghouse WL759) in a circuit developed by Dr. R. J. Cashman.[17] The current in the crystals (active area 0.5×6 mm^2) charges a well-insulated condensor. As soon as the potential across the condensor reaches a defined value, the trigger tube starts, and the resultant impulse is counted by a suitable amplifier and an electromagnetic counting device. The frequency of this periodic discharging is a scale for the intensity of the impinging radiation and currents of the order of 10^{-11} amp. can easily be measured.

The copper target x-ray tube, used in a full wave rectifier circuit, was applied with reduced heating current. The number of quanta emitted by the tube were determined by a Geiger-Müller counter with a measured efficiency of 40 percent for the copper K_α line monochromatized by a simple rocksalt spectrometer. The exit slit in front of the Geiger-Müller tube of 3 mm^2 defined the same bundle of x-rays with the crystal and with the counter.

With 1900 quanta hitting the crystal the condensor was charged within 12 to 13 sec. to the starting potential of the trigger tube. With this irradiation (about 150 quanta per sec.) a periodical trigging every 12 to 13 sec. is produced. This state is reached three minutes after the irradiation of the crystal has started. However, 20 min. after the x-ray irradiation was interrupted the crystal was still counting with a rate of one triggering in 55 to 59 sec. This persistent conductivity can be quickly reduced to zero, if a small amount of infra-red simultaneously illuminates the crystal during the irradiation with x-rays. The conductivity reached under x-ray irradiation is lowered somewhat by the simultaneous infra-red irradiation. 2300 quanta are necessary to charge the smallest capacity so that the trigger tube starts every 11 sec. Under the additional infra-red irradiation the trigger tube stops after counting two or three more counts (equivalent to 2300 quanta each), as soon as the x-ray irradiation is interrupted. No more counts were observed in the following 10 hr. The additional infra-red irradiation can also be applied between the different triggerings. It should be possible to design a circuit where each triggering flashes a small infra-red lamp (for instance, a xenon discharge), which reduces the persistent conductivity.

Thus, the trigger circuit is quite promising for the measuring of small numbers of x-ray quanta. Further experiments have to be made to find the actual limit of this method. As the current at 150 x-ray quanta per sec. is of the order of 1×10^{-11} amp. $=6.3\times10^7$ electron charges per sec. the efficiency in this case is 4.2×10^5 electrons per x-ray quantum.

CONCLUSIONS

The measurements show that the sensitivity of the CdS crystal can be compared directly with the sensitivity of the photo-multiplier x-ray detector. The crystal has in its present form the disadvantage that the response to x-ray pulses is apparently slower and that the current reaches a final value only after long irradiation. As the persistence in the original zinc sulfide screens has been greatly reduced by "killers," it might be possible to reduce the inertial of CdS by suitable impurities. On the other hand, the crystals do not show the fatigue effects observed with multiplier tubes. Their small size is a further advantage in some applications. As the crystal size defines the sensitive volume, it is not necessary to screen the receiver with lead slits.

The crystals can be grown in large numbers. They afford no treatment but a suitable mounting, for instance on a polystyrene strip, and the application of the Aquadag electrodes. As the crystal current is proportional to the first power of the potential applied no power of supply of extreme constancy is necessary.

The present measurements deal with such intensities that the effects of single x-ray quanta have not been considered. The regularities of electron pulses produced in these crystals afford further detailed investigations.[18]

[17] W. B. Nottingham, Rev. Sci. Inst. 11, 2 (1940). R. J. Cashman, reference 10, p. 918, Fig. 20.

[18] S. G. Zizzo and J. B. Platt, Phys. Rev. 76, 704 (1949) studied the pulses due to x-ray quanta as well as the continuous photocurrent in CdS crystals obtained from the author.

Note added in proof: Measurements of the efficiency of the photocurrents produced in CdS crystals by irradiation with x-rays have been recently published by Y. Fassbender and O. Hachenberg, Ann. d. Physik (6) 6, 229 (1949). They confirm our results regarding the amplification factor in CdS.

Diffraction by a Cylindrical Obstacle*

CHARLES H. PAPAS
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received September 23, 1949)

The diffraction of a plane electromagnetic wave by an infinitely long, perfectly conducting cylinder has been treated by a variational method (see the two papers by H. Levine and J. Schwinger). The incident field is assumed to be polarized in the direction of the cylinder axis, and thus the entire field is of two-dimensional nature. This formulation yields an expression for the diffracted cylindrical wave amplitude, at distances from the cylinder large compared to its transverse dimension and the wave-length, which is stationary relative to small independent variations of the surface currents arising from plane-wave excitation along a pair of directions in space; furthermore, the stationary form of the diffracted amplitude is independent of the scale of the surface currents. In accordance with a theorem of Levine and Schwinger, the total plane-wave scattering cross section is simply related to the diffracted cylindrical wave amplitude in the direction of incidence. To examine the high frequency behavior of the cross section, the surface current induced by a plane wave is taken different from zero only on the illuminated part of the cylinder, where its value is derived from the tangential component of the incident magnetic field. The resulting cross section is obtained and is shown to approach $4a$ when ka approaches infinity ($k=2\pi\div$wave length, a equals the radius of cylinder).

I. INTRODUCTION

THE theoretical examination of the diffraction of a steady-state plane electromagnetic wave by a cylindrical obstacle has interested many investigators. One of the earliest treatments proceeds according to the Fourier-Lamé eigenfunction method. The incident electromagnetic wave is assumed to be polarized parallel to the axis of the perfectly conducting, circular cylinder. The axis of the cylinder lies along the x axis, and the incident wave has only an x component of the electric field, $E_x^{\text{inc}}(\rho,\phi)$, and a y component of the magnetic field, $H_y^{\text{inc}}(\rho,\phi)$ (Fig. 1). For an incident wave whose electric field is of unit amplitude and propagating in the direction ϕ_1, i.e., $E_x^{\text{inc}}=\exp[ik\rho\cos(\phi-\phi_1)]$, the corresponding scattered electric field is given by

$$E_x^{sc}(\rho,\phi-\phi_1)=-\sum_{-\infty}^{\infty}(i)^m\frac{J_m(ka)}{H_m^{(1)}(ka)}H_m^{(1)}(k\rho)\cos m(\phi-\phi_1), \quad (1)$$

where $J_m(x)$ are the Bessel functions, and $H_m^{(1)}(x)$ are the Hankel functions of the first kind, and $k=2\pi\div$wave-length.

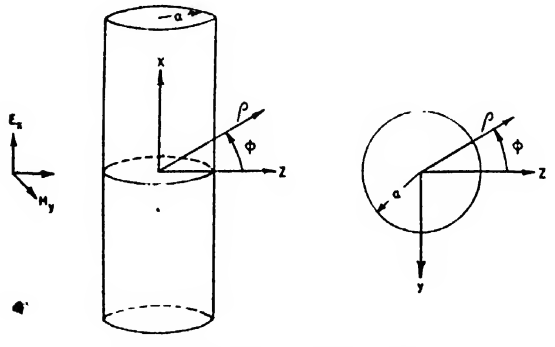


FIG. 1. Diffracting circular cylinder.

* The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (ONR), the Signal Corps of the U. S. Army, and the U. S. Air Force, under ONR Contract N5ori-76, T. O. I.

The inadequacy of the rigorous representation (1) stems from the fact that its convergence becomes slower with increasing values of ka, i.e., at high frequencies. To avoid this difficulty, Debye[1] represented the scattered field by a Fourier integral in the following manner. Let

$$f(m)=\frac{H_m^{(1)}(k\rho)}{H_m^{(1)}(ka)}\cos m(\phi-\phi_1),$$

where m can be considered to be a continuous real variable. According to the Fourier integral theorem,

$$f(m)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\mu\int_{\infty}^{\infty}f(\alpha)\exp[i\mu(m-\alpha)]d\alpha$$

$$=\frac{1}{\pi}\int_{-\infty}^{\infty}d\mu\exp(i\mu m)\int_0^{\infty}f(\alpha)\cdot\cos\mu\alpha\cdot d\alpha$$

$$=\frac{1}{\pi}\int_{-\infty}^{\infty}d\mu\exp(i\mu m)\int_0^{\infty}\frac{H_\alpha^{(1)}(k\rho)}{H_\alpha^{(1)}(ka)}\cdot\cos\alpha(\phi-\phi_1)\cdot\cos\mu\alpha\cdot d\alpha.$$

Then,

$$E_x^{sc}(\rho,\phi-\phi_1)=\frac{1}{\pi}\sum_{m=-\infty}^{\infty}(i)^mJ_m(ka)\cdot\int_{-\infty}^{\infty}d\mu\exp(i\mu m)\int_0^{\infty}\frac{H_\alpha^{(1)}(k\rho)}{H_\alpha^{(1)}(ka)}\cos\alpha(\phi-\phi_1)\cdot\cos\mu\alpha\cdot d\alpha.$$

Since

$$\sum_{m=-\infty}^{\infty}(i)^mJ_m(ka)\exp(im\mu)=\exp(ika\cos\mu),$$

by interchanging the order of summation and integra-

[1] P. Debye, "Das electromagnetische Feld um einem Zylinder und die Theorie des Regenbogens," Physik. Zeits. 9, 775–778 (1908).

tion Debye's integral representation is obtained:

$$E_x^{sc}(\rho,\phi-\phi_1)=-\frac{1}{\pi}\int_{-\infty}^{\infty}\exp(ika\cos\mu)d\mu \times\int_0^{\infty}\frac{H_\alpha^{(1)}(k\rho)}{H_\alpha^{(1)}(ka)}\cos\alpha(\phi-\phi_1)\cdot\cos\mu\alpha\cdot d\alpha. \quad (2)$$

By an approximate evaluation of (2) for large ka, he obtained

$$E_x^{sc}(\rho,\phi)=-\left(\frac{ka\cos(\phi/2)}{2k\rho}\right)^{\frac{1}{2}}\exp\left[ik\left(\rho-2a\cos\frac{\phi}{2}\right)\right] \quad (3)$$

for the scattered electric field resulting from a plane wave incident in the direction $\phi_1=0$.

Up to the present we have had the rigorous eigenfunction expansion of the scattered field (1) which suffers from convergence difficulties for large ka, and the Debye summation scheme yielding the expression (3) for the high frequency far-zone scattered field, valid for directions not in the neighborhood of $\phi=\pi$. Consequently (3) cannot be used to calculate the total scattering cross section of the cylinder; (1) can be used, however, even for large values of ka, i.e., at high frequencies, but only with considerable heuristic manipulation of the slowly convergent eigenfunction expansion.

It is the purpose of this paper to show that the aforementioned difficulties can be avoided if the problem is formulated from the start using a variational principle in the manner of Levine and Schwinger.[2,3] This formulation yields an expression for the diffracted cylindrical wave amplitude, at distances from the cylinder large compared to its transverse dimension and the wavelength, which is stationary relative to small independent variations of the surface currents arising from plane-wave excitation along a pair of directions in space. Furthermore, the stationary form of the diffracted amplitude is independent of the scale of the surface currents. The total plane-wave scattering cross section is simply related to the diffracted cylindrical wave amplitude in the direction of incidence. To examine the high frequency behavior of the cross section ($ka\gg1$), the surface current induced by a plane wave is taken different from zero only on the illuminated part of the cylinder, where its value is derived from the tengential component of the incident magnetic field.

II. INTEGRAL EQUATION FORMULATION

Consider a plane electromagnetic wave incident upon a perfectly conducting circular cylinder of radius a, whose axis lies along the x axis of a rectangular coordinate system, Fig. 1. The incident wave is polarized parallel to the axis of this cylinder, its electric field

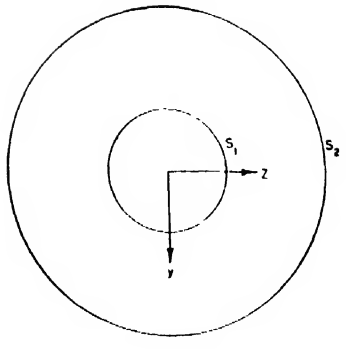


FIG. 2. Integration domain.

being given by

$$E_x^{inc}(\rho,\phi)=\exp[ik\rho\cos(\phi-\phi_1)], \quad (1)$$

where ρ,ϕ are cylindrical coordinates, $k=2\pi\div$ wavelength, and ϕ_1 is the direction of propagation. The scattered electric field, $E_x^{sc}(\rho,\phi)$, is also polarized parallel to the cylinder axis. The electric field at any point r, r being a two-dimensional vector from the origin to any point (ρ,ϕ), is the sum of the incident and scattered fields, $E_x(r)=E_x^{inc}(r)+E_x^{sc}(r)$. The problem is formulated in terms of these electric fields. To emphasize their two-dimensional scalar nature, denote them by $\psi(r)\equiv E_x(r)$, $\psi^{inc}(r)\equiv E_x^{inc}(r)$, and $\psi^{sc}(r)\equiv E_x^{sc}(r)$.

Note that $\psi(r)$ satisfies the scalar wave equation

$$(\nabla^2+k^2)\psi(r)=0, \quad (2)$$

and the two-dimensional scalar Green's function $G(r,r')$ satisfies the inhomogeneous scalar wave equation

$$(\nabla^2+k^2)G(r,r')=-\delta(r-r'), \quad (3)$$

where the inhomogeneous term is the two-dimensional Dirac delta-function. Applying Green's second scalar identity to $\psi(r)$ and $G(r,r')$, we obtain

$$\int_v[G(r',r)(\nabla'^2+k^2)\psi(r')-\psi(r')(\nabla'^2+k^2)G(r',r)]d\tau = \int_{s_1+s_2}\left[G(r',r)\frac{\partial}{\partial n'}\psi(r')-\psi(r')\frac{\partial}{\partial n'}G(r',r)\right]ds', \quad (4)$$

where the derivative in the surface integral is taken in the direction of the outward normal (Fig. 2). The surface integral extends over the surface of the scattering cylinder, s_1, and a concentric cylinder, s_2, of very large radius. In view of (2) and (3) the volume integral becomes $\psi(r)$, and since $\psi(r)=0$ when r lies on the scattering cylinder, it follows that

$$\psi(r)=\int_{s_1}G(r',r)\frac{\partial}{\partial n'}\psi(r')ds' + \int_{s_2}\left[G(r',r)\frac{\partial}{\partial n'}\psi(r')-\psi(r')\frac{\partial}{\partial n'}G(r',r)\right]ds'. \quad (5)$$

Recalling that $\psi(r)=\psi^{inc}(r)+\psi^{sc}(r)$, the integral over s_2

[2] H. Levine and J. Schwinger, "On the theory of diffraction by an aperture in an infinite plane screen, I," Phys. Rev. 74, 958–974 (1948).

[3] H. Levine and J. Schwinger, "On the theory of diffraction by an aperture in an infinite plane screen, II," Phys. Rev. 75, 1423–1432 (1949).

in (5) becomes

$$\int_{s_0}\left[G(r',r)\frac{\partial}{\partial n'}\psi^{\mathrm{inc}}(r')-\psi^{\mathrm{inc}}(r')\frac{\partial}{\partial n'}G(r',r)\right]ds'$$

$$+\int_{s_2}\left[G(r',r)\frac{\partial}{\partial n'}\psi^{s}(r')-\psi^{s}(r')\frac{\partial}{\partial n'}G(r',r)\right]ds'.$$

The first integral yields $\psi^{\mathrm{inc}}(r)$ and the second disappears due to the asymptotic behavior of the integrand for $r'\to\infty$, and (5) takes the form

$$\psi(r)=\psi^{\mathrm{inc}}(r)+\int_{s_1}G(r',r)\frac{\partial}{\partial n'}\psi(r')ds'. \tag{6}$$

Since $(\partial/\partial n')\psi(r')=-(\partial/\partial\rho')E_z(\rho',\phi')=i\omega\mu H_\phi(\rho',\phi')$ and $ds'=ad\phi'$, (6) becomes

$$\psi(\rho,\phi)=\psi^{\mathrm{inc}}(\rho,\phi)+\int_0^{2\pi}i\omega\mu H_\phi(a,\phi')G(a,\phi';\rho,\phi)\cdot ad\phi'. \tag{7}$$

$H_\phi(a,\phi')$ is numerically equal to the surface-current density at the point (a,ϕ') on the cylinder. Let $I_{\phi_1}(\phi')=i\omega\mu a H_\phi(a,\phi')$, where $I_{\phi_1}(\phi')$ is proportional to the current density at the point (a,ϕ') due to the incident wave (1) propagating in the direction ϕ_1. With this notation, (7) becomes

$$\psi(\rho,\phi)=\psi^{\mathrm{inc}}(\rho,\phi)+\int_0^{2\pi}I_{\phi_1}(\phi')G(a,\phi';\rho,\phi)d\phi'. \tag{8}$$

Since $\psi(a,\phi)=0$, i.e., E_z must disappear at the surface of the cylinder, the following integral equation is obtained by placing $\rho=a$ in (8):

$$0=\psi^{\mathrm{inc}}(\rho,\phi)+\int_0^{2\pi}I_{\phi_1}(\phi')G(a,\phi';a,\phi)d\phi'$$

or

$$-\exp[ika\cos(\phi-\phi_1)]=\int_0^{2\pi}I_{\phi_1}(\phi')G(a,\phi';a,\phi)d\phi' \tag{9}$$

The solution of this integral equation would yield $I_{\phi_1}(\phi)$, i.e., the current except for a numerical factor, produced by a plane incident wave traveling in the direction ϕ_1. By substituting this solution back into (8) and carrying out the indicated integration, the diffracted field $\psi(\rho,\phi)$ could be determined. Two objections are immediately apparent, the first is the difficult problem of solving (9) for $I_{\phi_1}(\phi')$, the second is the evaluation of the integral in (8) once $I_{\phi_1}(\phi')$ is determined. Moreover, to obtain the total scattering cross section, still another integration would have to be performed. Use will be made of the integral equation (9) to formulate a variational principle for the scattered wave amplitude.

III. VARIATIONAL PRINCIPLE FOR SCATTERED WAVE AMPLITUDE

Since the two-dimensional Green's function, $G(r,r')$, has to satisfy the inhomogeneous wave equation (3) and the radiation condition, i.e., it must represent at large distances from the cylinder outwardly propagating cylindrical waves, it follows that $G(r,r')=(i/4)H_0\times(k|r-r'|)$. For large r, the leading term of the asymptotic expansion is

$$r\to\infty:\quad G(r,r')\sim\frac{i}{4}\left(\frac{2}{\pi k|r-r'|}\right)^{\frac{1}{2}}\exp(ik)|r-r'|\exp\left(-i\frac{\pi}{4}\right),$$

where

$$|r-r'|=[\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')]\sim\rho\left(1-\frac{\rho'}{\rho}\cos(\phi-\phi')\right).$$

With these asymptotic values the scattered field, which is given by the integral in (8), at large distances becomes

$$\psi^{sc}(\rho,\phi)\sim\frac{i}{4}\left(\frac{2}{\pi ik\rho}\right)^{\frac{1}{2}}\exp(ik\rho)\times\left[\int_0^{2\pi}\exp[-ika\cos(\phi-\phi')]I_{\phi_1}(\phi')d\phi'\right].$$

The scattered wave amplitude is represented by the quantity in square brackets. We denote it by $A(\phi,\phi_1)$:

$$A(\phi,\phi_1)=\int_0^{2\pi}\exp[-ika\cos(\phi-\phi')]I_{\phi_1}(\phi')d\phi'. \tag{10}$$

As the notation explicitly denotes, $A(\phi,\phi_1)$ is the scattered wave amplitude in the direction ϕ for an incident wave traveling in the direction ϕ_1.

Multiplication of the integral equation (9) by $I_{\phi_2}(\phi)$ and integration with respect to ϕ (in accordance with the notation that $I_{\phi_2}(\phi)$ is $i\omega\mu a$ times the surface-current density at the point (a,ϕ) due to an incident plane wave traveling in the direction ϕ_2) leads to

$$-\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi=\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(a,\phi';a,\phi)I_{\phi_1}(\phi')d\phi'. \tag{11}$$

But according to (10),

$$\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi=\int_0^{2\pi}\exp[-ika\cos(\pi+\phi_1-\phi)]I_{\phi_2}(\phi)d\phi=A(\pi+\phi_1,\phi_2). \tag{12}$$

$A(\pi+\phi_1,\phi_2)$ represents the scattered wave amplitude in the direction of $\pi+\phi_1$ due to an incident wave traveling in the direction ϕ_2, and it is clear that it must equal the scattered wave amplitude in the direction $\pi+\phi_2$ due to an incident wave traveling in the direction ϕ_1. This reciprocity relation is given by the equality,

$$A(\pi+\phi_1,\phi_2)=A(\pi+\phi_2,\phi_1). \tag{13}$$

Dividing (11) by

$$\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi \times\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]I_{\phi_1}(\phi)d\phi,$$

and taking into consideration (12) and (13),

$$\frac{1}{A(\pi+\phi_2,\phi_1)}=\frac{1}{A(\pi+\phi_1,\phi_2)}\frac{\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(a,\phi';a,\phi)I_{\phi_1}(\phi')d\phi'}{\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]I_{\phi_1}(\phi)d\phi} \tag{14}$$

The expression (14) is homogeneous in $I_{\phi_1}(\phi)$, $I_{\phi_2}(\phi)$ and stationary with respect to independent first-order variations about their correct values (determined by integral equations of the form (9)). The property of homogeneity is clear. To show the stationary property of (14), proceed as follows. Clearing fractions in (14),

$$\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi \times\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]I_{\phi_1}(\phi)d\phi = -A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi',$$

where $A\equiv A(\pi+\phi_1,\phi_2)$, $G(\phi',\phi)\equiv G(a,\phi';a,\phi)$. Now take the variations of $I_{\phi_2}(\phi)$, $I_{\phi_1}(\phi)$, and A.

$$\begin{aligned}&\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]\delta I_{\phi_2}(\phi)d\phi \times\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]I_{\phi_1}(\phi)d\phi\\ &+\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]I_{\phi_2}(\phi)d\phi \times\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]\delta I_{\phi_1}(\phi)d\phi_1\\ &=-\delta A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi'\\ &-A\int_0^{2\pi}\int_0^{2\pi}d\phi\delta I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi'\\ &-A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')\delta I_{\phi_1}(\phi')d\phi'.\end{aligned}$$

Using (12) and (13) this can be written in the form

$$\begin{aligned}&A\int_0^{2\pi}\exp[ika\cos(\phi-\phi_1)]\delta I_{\phi_2}(\phi)d\phi\\ &+A\int_0^{2\pi}\exp[ika\cos(\phi-\phi_2)]\delta I_{\phi_1}(\phi)d\phi\\ &=-\delta A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi'\\ &-A\int_0^{2\pi}\int_0^{2\pi}d\phi\delta I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi'\\ &-A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')\delta I_{\phi_1}(\phi')d\phi'.\end{aligned}$$

Combining terms, this equation becomes

$$\begin{aligned}&\delta A\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\phi_2}(\phi)G(\phi,\phi')I_{\phi_1}(\phi')d\phi'\\ &=-A\int_0^{2\pi}d\phi\delta I_{\phi_2}(\phi)\Big[\int_0^{2\pi}d\phi'G(\phi,\phi')I_{\phi_1}(\phi')\\ &+\exp[ika\cos(\phi-\phi_1)]\Big]-A\int_0^{2\pi}d\phi'\delta I_{\phi_1}(\phi')\\ &\Big[\int_0^{2\pi}d\phi G(\phi,\phi')I_{\phi_2}(\phi)\\ &+\exp[ika\cos(\phi'-\phi_2)]\Big].\end{aligned} \tag{15}$$

If $I_{\phi_1}(\phi)$ and $I_{\phi_2}(\phi)$ satisfy integral equations of the form (9), then the square-bracket terms in (15) vanish identically. Hence A is stationary with respect to first-order independent variations of $I_{\phi_1}(\phi)$ and $I_{\phi_2}(\phi)$.

IV. A THEOREM CONCERNING THE TOTAL SCATTERING CROSS SECTION AND SCATTERED WAVE AMPLITUDE

The total scattering cross section is defined by the ratio of the total scattered power per unit length of

cylinder to the incident power per unit area. The time-average incident power per unit area, $\bar{S}^{\text{inc}}$, is given by $\frac{1}{2}\,\text{Re}E\times H^*=\frac{1}{2}\,\text{Re}E^*\times H$, where the asterisk indicates the conjugate complex and Re the real part. Since

$$E_x^{\text{inc}}=\exp(ikz)$$

and

$$H_y^{\text{inc}}=\frac{1}{i\omega\mu}\frac{\partial}{\partial z}E_x^{\text{inc}}=\frac{k}{\omega\mu}\exp(ikz),$$

then

$$\bar{S}^{\text{inc}}=\tfrac{1}{2}\,\text{Re}E_x^{*\text{inc}}H_y^{\text{inc}}=\frac{k}{2\omega\mu}.$$

The time-average scattered power, $\bar{P}^{sc}$, passing through a cylinder s_2 of large radius is equal to the time-average power leaving the cylindrical surface s_1, i.e.,

$$\bar{P}^{sc}=\text{Re}\int_{s_2}\tfrac{1}{2}(E_x^{sc}(\rho,\phi))(-H_\phi^{sc}(\rho,\phi))\rho d\phi$$

$$=\text{Re}\int_{s_1}\tfrac{1}{2}(E_x^{sc}(a,\phi))(-H_\phi^{sc}(a,\phi))ad\phi.$$

The total scattering cross section, σ, is then given by the expression

$$\sigma=\frac{\bar{P}^{sc}}{\bar{S}^{\text{inc}}}=\frac{1}{k}\,\text{Re}\frac{1}{i}\int_0^{2\pi}E_x^{*sc}(a,\phi)\frac{\partial}{\partial\rho}E_x^{sc}(a,\phi)ad\phi. \quad (16)$$

At the surface of the cylinder, $E_x^{*sc}(a,\phi)=-E_x^{*\text{inc}}(a,\phi)$; hence

$$\sigma=\frac{1}{k}\,\text{Re}\frac{1}{i}\int_0^{2\pi}-E_x^{*\text{inc}}(a,\phi)\frac{\partial}{\partial\rho}[E_x(a,\phi)-E_x^{\text{inc}}(a,\phi)]ad\phi.$$

But

$$\int_0^{2\pi}E_x^{*\text{inc}}(a,\phi)\frac{\partial}{\partial\rho}E_x^{\text{inc}}(a,\phi)ad\phi=0,$$

and consequently

$$\sigma=\frac{1}{k}\,\text{Re}\frac{1}{i}\int_0^{2\pi}\exp[-ika\cos(\phi-\phi_1)]I_{\phi_1}(\phi)d\phi, \quad (17)$$

where use has been made of $a(\partial/\partial\rho)E_x(a,\phi)=-i\omega\mu aH_\phi\times(a,\phi)=-I_{\phi_1}(\phi)$. In view of (10) and (17) σ can be represented in terms of the scattered wave amplitude:

$$\sigma=\frac{1}{k}\,\text{Re}\frac{1}{i}A(\phi_1,\phi_1)=\frac{1}{k}\,\text{Im}A(\phi_1,\phi_1), \quad (18)$$

where Im means imaginary part.

V. LOW FREQUENCY SCATTERING CROSS SECTION

To find the total scattering cross section it is necessary first to compute $A(\phi_1,\phi_1)$ from (14) and then to use (18). Let $\phi_1=0$ and $\phi_2=\pi$ in (14); then

$$\frac{1}{A(\pi,\pi)}=\frac{\int_0^{2\pi}\int_0^{2\pi}d\phi I_\pi(\phi)G(a,\phi';a,\phi)I_0(\phi')d\phi'}{\int_0^{2\pi}\exp(ika\cos\phi)I_\pi(\phi)d\phi\int_0^{2\pi}\exp(-ika\cos\phi)I_0(\phi)d\phi}. \quad (19)$$

At low frequencies $ka\ll1$; also choose $I_\pi(\phi)=I_0(\phi)=1$. The numerator then becomes

$$\int_0^{2\pi}\int_0^{2\pi}d\phi G(a,\phi';a,\phi)d\phi'=\frac{i}{4}\int_0^{2\pi}\int_0^{2\pi}d\phi d\phi' H_0(k|r-r'|)$$

$$=\frac{i}{4}\int_0^{2\pi}\int_0^{2\pi}d\phi d\phi'\sum_{m=0}'\epsilon_m J_m(ka)H_m^{(1)}(ka)\cos m(\phi-\phi')$$

$$=i\pi^2J_0(ka)H_0(ka).$$

and

$$\int_0^{2\pi}\exp(ika\cos\phi)d\phi=2\pi J_0(ka),$$

$$\int_0^{2\pi}\exp(-ika\cos\phi)d\phi=2\pi J_0(ka).$$

Consequently

$$\frac{1}{A(\pi,\pi)}=\frac{i\pi^2J_0(ka)H_0(ka)}{[2\pi J_0(ka)]^2}=\frac{i}{4}\frac{H_0(ka)}{J_0(ka)},$$

and

$$\sigma=\frac{4}{k}\frac{J_0^2(ka)}{J_0^2(ka)+N_0^2(ka)}. \quad (20)$$

Suppose, instead of taking $I_\pi(\phi)=I_0(\phi)=1$, the following more general representation is used:

$$I_0(\phi)=\sum_0^\infty A_m\cos m\phi,$$

$$I_\pi(\phi)=\sum_0^\infty A_m\cos m(\phi-\pi)=\sum_0^\infty A_m(-1)^m\cos m\phi,$$

where the A_m's are constants. Substitution in (19) gives for the numerator:

$$\int_0^{2\pi}\int_0^{2\pi}d\phi I_\pi(\phi)G(a,\phi';a,\phi)I_0(\phi')d\phi'$$

$$=\frac{i}{4}\int_0^{2\pi}\int_0^{2\pi}d\phi d\phi'\left(\sum_{m=0}^\infty A_m(-1)^m\cos m\phi\right)$$

$$\times\left(\sum_{p=0}^\infty\epsilon_pJ_p(ka)H_p^{(1)}(ka)\cos p(\phi-\phi')\right)$$

$$\times\left(\sum_{n=0}^\infty A_n\cos n\phi'\right).$$

Let

$$C_{mn}=(-1)^m\frac{i}{4}\int_0^{2\pi}\int_0^{2\pi}d\phi d\phi'\cos m\phi \times\left(\sum_{p=0}^{\infty}\epsilon_p J_p(ka)H_p^{(1)}(ka)\cos p(\phi-\phi')\right)\cos n\phi'.$$

But

$$\int_0^{2\pi}\int_0^{2\pi}d\phi d\phi'\cos m\phi\cos p(\phi-\phi')\cos n\phi'=\frac{4\pi^2\delta_p{}^n\delta_p{}^m}{\epsilon_p{}^2},$$

where

$$\delta_n{}^m=\begin{cases}1 & m=n\\ 0 & m\neq n\end{cases}\quad\text{and}\quad\epsilon_m=\begin{cases}1 & m=0\\ 2 & m\geq 1\end{cases},$$

and hence

$$C_{mn}=\frac{(-1)^m i\pi^2}{(\epsilon_m)}\delta_n{}^m J_m(ka)H_m^{(1)}(ka).$$

The numerator becomes the double series,

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}A_m C_{mn}A_n.$$

For the denominator the following is obtained:

$$\int_0^{2\pi}\exp(ika\cos\phi)I_\pi(\phi)d\phi\int_0^{2\pi}\exp(-ika\cos\phi)I_0(\phi)d\phi$$

$$=\int_0^{2\pi}\exp(ika\cos)\phi\sum_{m=0}^{\infty}A_m(-1)^m\cos m\phi d\phi \times\int_0^{2\pi}\exp(-ika\cos)\phi\sum_{n=0}^{\infty}A_n\cos n\phi d\phi$$

$$=\left(\sum_{m=0}^{\infty}A_m(-1)^m 2\pi(i)^m J_m(ka)\right)\times\left(\sum_{n=0}^{\infty}A_n 2\pi(-i)^m J_n(ka)\right).$$

By setting $B_m=2\pi(-i)^m J_m(ka)$, it follows that

$$\frac{1}{A(\pi,\pi)}=\frac{\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}A_m C_{mn}A_n}{\left(\sum_{l=0}^{\infty}A_l B_l\right)^2}. \tag{21}$$

In view of the stationary character of $A(\pi,\pi)$, differentiation of (21) with respect to A_m gives

$$A(\pi,\pi)\sum_{n=0}^{\infty}A_n C_{mn}=-B_m\sum_{n=0}^{\infty}A_n B_n,\quad m=0,1,2,\cdots \tag{22}$$

To simplify (22) the coefficients D_n are introduced so that

$$A(\pi,\pi)=-\sum_{n=0}^{\infty}B_n D_n, \tag{23}$$

where the D_n's are determined by the simultaneous equations

$$\sum_{n=0}^{\infty}C_{mn}D_n=B_m,\quad m=0,1,2,\cdots \tag{24}$$

Since $C_{mn}=0$ when $m\neq n$ only the diagonal terms of (24) are different from zero, i.e., $D_n=B_n/C_{nn}$, and hence

$$A(\pi,\pi)=-\sum_{n=0}^{\infty}\frac{B_n{}^2}{C_{nn}}=\sum_{n=0}^{\infty}\frac{(\epsilon_n 2\pi(-i)^n J_n(ka))^2}{i\pi^2(-1)^n J_n(ka)H_n^{(1)}(ka)}=4i\sum_{n=0}^{\infty}\epsilon_n\frac{J_n(ka)}{H_n^{(1)}(ka)}. \tag{25}$$

Applying the theorem (18) to (25) the total scattering cross section is obtained. It is

$$\sigma=\frac{4}{k}\sum_{n=0}^{\infty}\frac{\epsilon_n J_n{}^2(ka)}{J_n{}^2(ka)+N_n{}^2(ka)}, \tag{26}$$

where use has been made of the form $H_n^{(1)}(ka)=J_n(ka)+iN_n(ka)$. Since $ka\ll 1$, only the first term of (26) contributes appreciably to the value of σ. Therefore

$$\sigma=\frac{4}{k}\frac{J_0{}^2(ka)}{J_0{}^2(ka)+N_0{}^2(ka)}. \tag{27}$$

Comparing (20) with (27) it is seen that the expressions are identical. It appears that for low frequencies ($ka\ll 1$) very accurate results are obtained by taking the current on the cylinder to be independent of ϕ.

VI. HIGH FREQUENCY SCATTERING CROSS SECTION

Now consider the frequency of the incident wave to be high, i.e., $ka\gg 1$. In (14) let $\phi_1=3\pi/2$, $\phi_2=\pi/2$, and for this choice the following result is obtained:

$$\frac{1}{A\left(\pi+\frac{3\pi}{2},\frac{\pi}{2}\right)}=\frac{\int_0^{2\pi}\int_0^{2\pi}d\phi I_{\pi/2}(\phi)G(a,\phi';a,\phi)I_{3\pi/2}(\phi')d\phi'}{\int_0^{2\pi}\exp\left[ika\cos\left(\phi-\frac{3\pi}{2}\right)\right]I_{\pi/2}(\phi)d\phi\int_0^{2\pi}\exp\left[ika\cos\left(\phi-\frac{\pi}{2}\right)\right]I_{3\pi/2}(\phi)d\phi}. \tag{28}$$

In accordance with the assumed convention, $I_{\pi/2}(\phi)$ and $I_{3\pi/2}(\phi)$ are proportional to the surface currents induced by incident waves propagating in the directions $\pi/2$ and $3\pi/2$ respectively. For an incident wave propagating in the direction ϕ_1, $H_\phi^{inc}(\rho,\phi)=-(k/\omega\mu)\times\cos(\phi-\phi_1)\exp[ik\rho\cos(\phi-\phi_1)]$. It is assumed that the high frequency current induced by a plane wave is different from zero only on the illuminated half of the cylinder, where its value is numerically equal to the tangential component of the incident magnetic field. That is, for a trial function we take

$$I_{\phi_1}(\phi)=\text{const.}\cos(\phi-\phi_1)\exp[ika\cos(\phi-\phi_1)]. \quad (29)$$

Hence,

$$I_{\pi/2}(\phi)=\text{const.}\sin\phi\exp[ika\sin\phi] \text{ for } \pi\leq\phi\leq 2\pi$$
$$=0 \text{ for } 0\leq\phi\leq\pi,$$

and

$$I_{3\pi/2}(\phi)=-\text{const.}\sin\phi\exp[-ika\sin\phi] \text{ for } 0\leq\phi\leq\pi$$
$$=0 \text{ for } \pi\leq\phi\leq 2\pi.$$

Substituting these trial functions in (28) and noting that $A(\pi+3\pi/2,\pi/2)=A(\pi/2,\pi/2)$, we get

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=\frac{\int_0^{\pi}d\phi'\sin\phi'\exp[-ika\sin\phi']\int_{\pi}^{2\pi}\sin\phi\exp(ika\sin\phi)G(a,\phi';a,\phi)}{\int_{\pi}^{2\pi}\sin\phi d\phi\int_0^{\pi}\sin\phi d\phi}$$

Since

$$G(a,\phi',a,\phi)=\frac{i}{4}H_0^{(1)}\left(2ka\left|\sin\left(\frac{\phi-\phi'}{2}\right)\right|\right),$$

then

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=\frac{i}{16}\int_0^{\pi}d\phi'\int_{\pi}^{2\pi}d\phi\sin\phi'\sin\phi\exp[-ika(\sin\phi'-\sin\phi)]H_0^{(1)}\left(2ka\left|\sin\left(\frac{\phi-\phi'}{2}\right)\right|\right). \quad (30)$$

Let $\phi=2\pi-\xi$. Then $\sin\phi=-\sin\xi$, $d\phi=-d\xi$, and

$$\left|\sin\left(\frac{\phi-\phi'}{2}\right)\right|=\left|\sin\left(\frac{\xi+\phi'}{2}\right)\right|.$$

With this change of variables (30) becomes*

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=-\frac{i}{16}\int_0^{\pi}d\phi'\int_0^{\pi}d\xi\sin\phi'\sin\xi$$
$$\times\exp[-ika(\sin\phi'+\sin\xi)]H_0^{(1)}\left(2ka\left|\sin\left(\frac{\xi+\phi'}{2}\right)\right|\right).$$

Let the variables be changed again according to the following transformation

$$(\phi'+\xi^2)/2=\beta+\pi/2, \quad (\phi'-\xi)/2=\alpha.$$

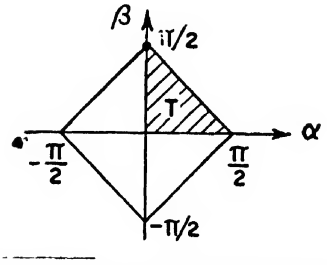


FIG. 3. Integration domain.

Then

$$\sin\phi'\sin\xi=\tfrac{1}{2}(\cos 2\alpha+\cos 2\beta), \quad \sin\phi'+\sin\xi=2\cos\alpha\cos\beta,$$
$$\frac{\partial(\phi',\xi)}{\partial(\alpha,\beta)}=2, \quad \text{and} \quad \left|\sin((\xi+\phi')/2)\right|=\cos\beta.$$

It follows that in terms of α and β,

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=-\frac{i}{16}\int\int_Q\exp(-2ika\cos\alpha\cos\beta)$$
$$\times(\cos^2\alpha+\cos 2\beta)H_0^{(1)}(2ka\cos\beta), \quad (31)$$

where Q is the square region in the α, β plane shown in Fig. 3. In view of the symmetry in α and β about $\alpha=\beta=0$, we can confine the integration to the triangular region T, Consequently,

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=-\frac{i}{4}\int_0^{\pi/2}d\beta\int_0^{\pi/2-\beta}d\alpha\exp(-2ika\cos\alpha\cos\beta)$$
$$\times(\cos 2\alpha+\cos 2\beta)H_0^{(1)}(2kR\cos\beta). \quad (32)$$

Consider first the integration with respect to α. Since $ka\gg 1$, the main contribution to the value of this integral occurs in the neighborhood of $\alpha=0$ (provided $\beta\neq\pi/2$). An approximate evaluation is obtained by

* The author received enlightening discussions on integrals of this class from Professor F. Tricomi.

the method of stationary phase:

$$\int_0^{\pi/2-\beta} d\alpha \exp(-2ika\cos\alpha\cos\beta)(\cos 2\alpha+\cos 2\beta)$$
$$=\left(\frac{\pi}{ka}\right)^{\frac{1}{2}}\frac{\cos^2\beta}{(\cos\beta)^{\frac{1}{2}}}\exp(-2ika\cos\beta)\exp\left(i\frac{\pi}{4}\right).$$

There remains the task of evaluating the following integral:

$$\frac{1}{A\left(\frac{\pi}{2},\frac{\pi}{2}\right)}=-\frac{i}{4}\left(\frac{\pi}{ka}\right)^{\frac{1}{2}}\exp\left(i\frac{\pi}{4}\right)\int_0^{\pi/2}d\beta\frac{\cos^2\beta}{(\cos\beta)^{\frac{1}{2}}}$$
$$\times\exp[-2ika\cos\beta]H_0^{(1)}(2ka\cos\beta). \quad (33)$$

As $\beta\to\pi/2$ the integrand disappears, and hence the contribution to the integral in the neighborhood of $\beta=\pi/2$ is very small. When $\beta\neq\pi/2$ and $ka\to\infty$,

$$H_0^{(1)}(2ka\cos\beta)$$
$$\sim\left(\frac{2}{2\pi ka\cos\beta}\right)^{\frac{1}{2}}\exp(2ika\cos\beta)\exp(-i\pi/4). \quad (34)$$

After substituting this asymptotic form into (33) and changing the upper limit from $\pi/2$ to β_0, the result is

$$\frac{1}{A(\pi/2,\pi/2)}\sim-\frac{i}{4}\frac{1}{ka}\int_0^{\beta_0}d\beta\cos\beta=-\frac{i}{4ka}\sin\beta_0 \quad (35)$$

for the leading term of the asymptotic development of $[1/A(\pi/2,\pi/2)]$ in reciprocal powers of ka. Application of the theorem (18) to (35) gives for the total scattering cross section

$$\sigma=\frac{1}{k}\mathrm{Im}A\left(\frac{\pi}{2},\frac{\pi}{2}\right)\sim\frac{4a}{\sin\beta_0}=\frac{4a}{(1-(4/ka)^2)^{\frac{1}{2}}}. \quad (36)$$

Choose β_0 such that $\cos\beta_0=4/ka$. This choice guarantees the argument of the Hankel function to be equal to or greater than 8 for the entire range of values from $\beta=0$ to $\beta=\beta_0$, and thereby allows the replacement of the Hankel function by the first term of its asymptotic development. It is seen that as $ka\to\infty$, $\sigma\to 4a$ which is the geometric optics scattering cross section and physically equal to twice the projected area of a unit length of the scattering cylinder. Figure 4 shows a plot of σk *versus* ka. For values of ka less than 10, Eq. (26) was used to compute σk; for ka larger than 8, Eq. (36) was used. In the region $8<ka<10$ the broken and solid curves overlap.

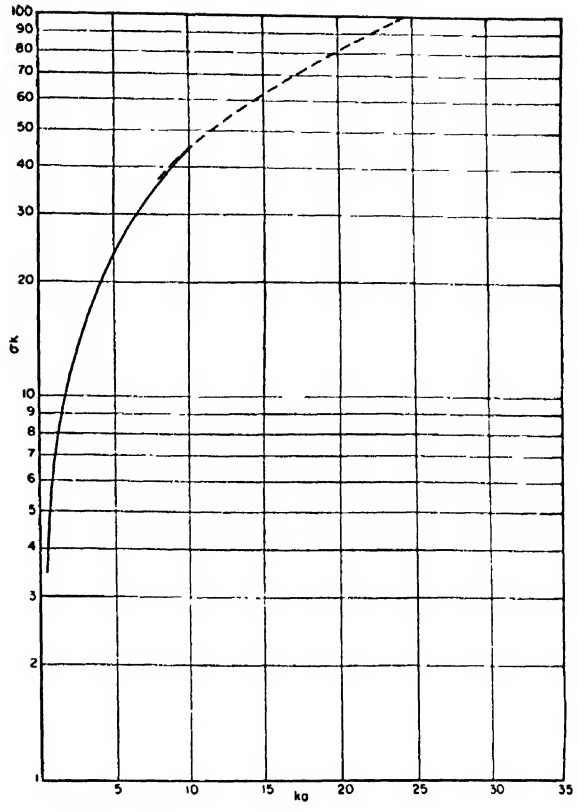


FIG. 4. σk *versus* ka. Solid curve computed fr
Broken curve computed from (36).

The author wishes to acknowledge the generous help of Dr. Harold Levine of Harvard University, without whose patient and clear explanations this paper would never have been completed. The assistance of Miss J. Klimas and Mrs. R. Stokey with numerical calculations is gratefully acknowledged.

Air Temperature Modification by Vertical Transport*

ROBERT G. FLEAGLE
University of Washington, Seattle, Washington
(Received October 3, 1949)

It is suggested that when the air temperature increases upward an efficient and economical method for modifying low temperatures near the ground is to force air downward through vertical stacks from the relatively warm air aloft. Theory is developed which shows that the necessary power per unit mass is mainly a function of the air velocity in the stack. The energy required by this method is compared with the energy required to heat the air originally near the ground, and it is shown that under a strong temperature inversion the power method requires considerably less energy than the heating method.

INTRODUCTION

TEMPERATURE modification by a combination of radiational cooling and direct application of heat (both accomplished by smudge pots) is the standard method used by fruit growers in avoiding low nocturnal temperatures near the ground. Under conditions requiring the use of smudge pots the temperature of the air usually increases rapidly with height; at 100–200 feet above the ground the temperature may be 5 to 10° C higher than the temperature near the ground. Recently, this normal temperature inversion has been utilized in modifying the temperatures near the ground; large fans have been set up above the trees to mix the cold and warm air.[1] It appears that a more effective and better controlled method would be to conduct the warm air downward through vertical stacks and exhaust it within the orchard. It is proposed to investigate the energy required for this process and to compare it with the energy required by direct heating. The application to maintenance of outdoor temperature sensitive installations, machinery, and supplies is obvious.

THEORY

Consider the energy required to force a unit mass of air through a vertical stack at a speed, v_2. Some energy is converted into kinetic energy, some is lost in friction at the sides of the stack, and some is required to overcome the buoyant force experienced by the warm air in being displaced downward into colder air. Where M represents the energy per gram supplied by the fan; F_r, the energy per gram dissipated by friction; and W, the energy per gram required to overcome atmospheric stability, the principle of conservation of energy requires that,

$$M=\tfrac{1}{2}v_2^2+F_r+W. \tag{1}$$

The mass of air passing through the stack per second is given by $\pi r^2\rho v_2$, where ρ represents air density and r represents radius of the stack. We therefore have for the power required of the motor,

$$P(\text{ergs/sec.})=\pi r^2\rho v_2\{\tfrac{1}{2}v_2^2+F_r+W\}. \tag{2}$$

The frictional loss of energy per unit mass in fundamental c.g.s. units is given by the empirical expression,

$$F_r=\frac{fhv_2^2}{4r}, \tag{3}$$

where f represents a dimensionless coefficient depending upon roughness of conduit and the Reynolds number and h represents length of the conduit.[2] Several values of f are shown in Table I below.[3] Table I shows that f is not sensitive to changes in r or v_2 within the ranges applicable to this problem. The single value of 0.02 will be used in future computations.

The work required to bring air down the stack against the buoyant force due to atmospheric stability remains to be computed. Where F represents the net upward

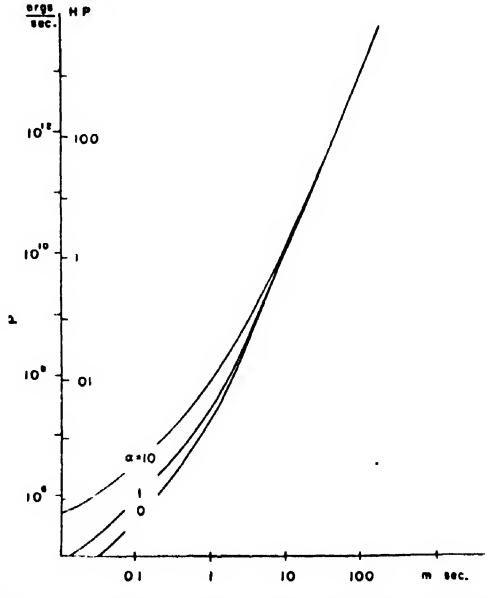


FIG. 1. Power required by fan as function of velocity for $\alpha=0$, 1, $10K/(100\ m)$ as computed from Eq. (12).

* Contribution No. 5, Department of Meteorology and Climatology, University of Washington.

[1] R. E. Ahrens, "Breath of steel," *Steelways*, No. 12, 32 (1947).

[2] Am. Soc. Heating and Ventilating Eng., *Heating, Ventilating, Air Conditioning Guide* (Waverly Press, Inc., Baltimore, Maryland, 1949), pp. 807–808.

[3] Computed from (3) using data supplied in Am. Soc. Heating and Ventilating Eng. (see reference 2).

force per gram on the displaced air, we have,

$$W=-\int_h^0 F\,dz. \tag{4}$$

The net force per gram is given by,

$$F=\frac{\rho-\rho'}{\rho'}g=\frac{(p/RT)-(p/RT')}{p/RT'}g=\frac{T'-T}{T}g, \tag{5}$$

where the prime indicates a property of the moving air, p represents pressure; T, absolute temperature; and R, the gas constant for air. Substitution of (5) in (4) gives,

$$W=-g\int_h^0 \frac{T'-T}{T}\,dz. \tag{6}$$

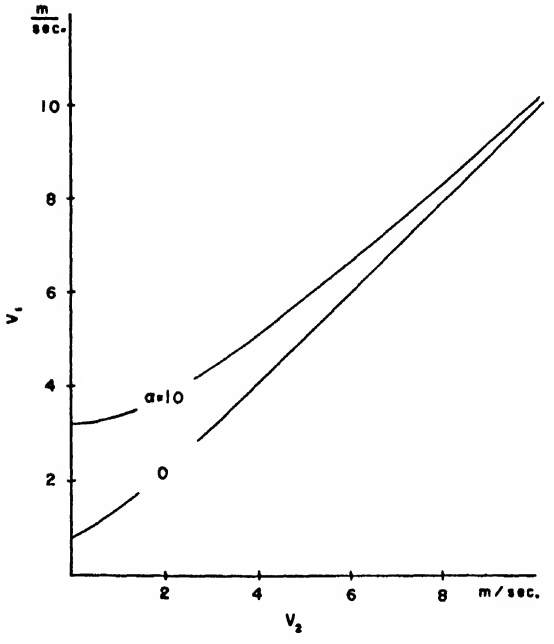


FIG. 2. Air speed within stack, (v_2), as a function of horizontal wind speed at the top of the stack, (v_1) for $\alpha=0$ and $10K/(100\text{ m})$ as computed from Eq. (13).

We shall assume that the temperature in the undisturbed air may be expressed by a linear function of elevation and that the disturbed air neither gains nor loses heat to the stack walls. Thus, the descending air increases in temperature at the adiabatic rate. We may express these assumptions by the following:

$$T=T_0+\alpha z \tag{7}$$

$$T'=T_0+\alpha h+\frac{g}{c_p}(h-z). \tag{8}$$

Equation (6) now may be written, for $\alpha\neq 0$

$$W=-\frac{g(\alpha+(g/c_p))}{\alpha}\int_h^0 \frac{(h-z)\partial z}{(T_0/\alpha)+z}. \tag{9}$$

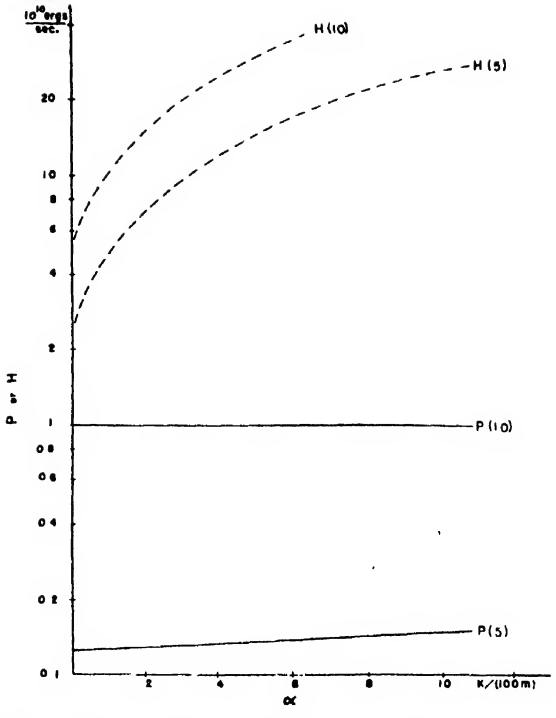


FIG. 3. Power required by fan compared with rate of energy required for heating (H) as function of rate of temperature increase with height (α). The bracketed numbers indicate the air speed within the stack in m/sec.

Integration gives,

$$W=-g\left(1+\frac{g}{\alpha c_p}\right)\left\{-\left(h+\frac{T_0}{\alpha}\right)\log_e\frac{T_h}{T_0}+h\right\}, \tag{10}$$

where T_h represents the temperature at the height of the top of the stack. For purposes of computation it is sufficiently accurate to replace $T_0+\alpha z$ by T_0 in the denominator of (9). This leads to,

$$W=gh^2/2T_0(\alpha+(g/c_p)). \tag{11}$$

Equation (11) holds for $\alpha=0$ since division by α is not required in the integration.

Finally, substituting (3) and (11) in (2) results in,

$$P=\pi r^2\rho v_2\left\{0.5v_2^2(1+0.01h/r)+\frac{gh^2}{2T_0}\left(\alpha+\frac{g}{c_p}\right)\right\}. \tag{12}$$

In order to visualize the relative importance of the individual terms in this rather cumbersome equation, P is plotted as a function of v_2 and α in Fig. 1 using the following arbitrary values:

$r=0.5$ meter
$h=50$ meters
$T_0=270°$ K

Figure 1 shows that within the power range between 0.1 and 100 horsepower the lapse rate does not affect seriously the air speed within the stack.

TABLE I.

r(cm)	v_1(m/sec.)	f
15	2.5	0.022
15	5.1	0.021
15	50.8	0.018
46	2.5	0.019
46	5.1	0.016
46	50.8	0.013

It is of interest to investigate the speed within the stack which is possible without the use of a fan. A horizontal opening at the top of the stack free to rotate so that it always faces the wind stream is assumed. For this case, the kinetic energy of the moving air must be added to the left side of (1) and M set equal to zero. Equation (12) then becomes,

$$0.5v_2^2\left(1+0.01\frac{h}{r}\right)+\frac{gh^2}{2T_0}\left(\alpha+\frac{g}{c_p}\right)=0.5v_1^2. \quad (13)$$

In Fig. 2, v_2 is plotted as a function of v_1 and α using the same arbitrary values listed above. Figure 2 indicates that the air speed within the stack is of the same order of magnitude as the wind speed. It appears that where only a small quantity of air is required, the kinetic energy of the wind at the height of the top of the stack might be sufficient to do the job alone. However, it should be realized that when the wind speed at the top of the stack exceeds about 5 m.p.sec. the resulting mechanical turbulence is likely to prevent the creation of a large temperature inversion. In other words, at high wind speeds the warm air is conducted to the ground without the assistance of fan and stack.

POWER COMPARISON

It is important to compare the energy required to produce at the ground a certain temperature increase by the method described here with the energy required to heat the original air through the same temperature interval. The rate of expenditure of heat required by the latter process may be written,

$$H=\pi r^2\rho v_2 c_p[T_h+(gh/c_p)-T_0]. \quad (14)$$

Substitution of $T_0+\alpha h$ for T_h leads to,

$$H=\pi r^2\rho v_2 c_p h(\alpha+(g/c_p)). \quad (15)$$

This energy requirement may be compared with P by plotting P and H as functions of α for certain arbitrary values of v_2, say 5 and 10 m/sec. In Fig. 3 P and H have been computed for air at sea level using the same values of r, h, and T_0 used before. Inspection of Fig. 3 shows that economy of the power method over the heating method ranges from a factor of about 5 to a factor of about 100. The relative economy varies in a direct sense with the strength of the temperature inversion. Economy of power is greater at the lower velocity (5 m/sec.) than at the higher velocity (10 m/sec.). It follows that maximum efficiency would be attained with fans operating at slow speed in stacks of large radius. Of course the economy in energy expended by the power method must be balanced against the greater cost of mechanical energy. Also, the greater efficiency attained with large stacks must be balanced against the greater initial cost of large stacks. Determination of the economy of cost of installation and operation depends upon local factors beyond the scope of this paper.

Aside from the question of economy, advantages of the power method over the heating method include: greater uniformity in temperature, absence of smoke, gas, and dirt, less attention required, and the possibility of utilizing the kinetic energy of the wind. Disadvantages include: more critical dependence on a large temperature inversion, greater initial expense (unless existing stacks may be used).

ACKNOWLEDGMENT

The writer is indebted to Mr. G. E. Ledbetter of the Department of Aeronautical Engineering for discussing with him the problem of frictional loss of energy.

Mode Conversion Losses in Transmission of Circular Electric Waves through Slightly Non-Cylindrical Guides

SAMUEL P. MORGAN, JR.
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received July 8, 1949)

A general expression is derived for the effective attenuation of circular electric (TE_{01}) waves owing to mode conversions in a section of wave guide whose shape deviates slightly in any specified manner from a perfect circular cylinder. Numerical predictions are in good agreement with experiment for the special case of transmission through an elliptically deformed section of pipe. The effect of random distortions in a long wave guide line is analyzed in terms of the mean-square values of the Fourier coefficients describing the distortions; and from the general formulas it appears that the mode conversion losses depend in large measure upon the statistical distribution of the various types of distortion. Under certain rather arbitrary simplifying assumptions about this distribution, it is calculated that mode conversions in a 4.732-inch copper pipe whose radius deviates by 1 mil r.m.s. from that of an average cylinder will increase the attenuation of the TE_{01} mode at 3.2 cm by an amount equal to 8 percent of the theoretical copper losses. The dependence on frequency of mode conversion losses in such a guide is discussed.

I. INTRODUCTION

THE propagation of electromagnetic energy through a perfectly conducting cylindrical guide may take place at sufficiently high frequencies in any one of a number of independent transmission modes, the field pattern for a given mode being the same, up to phase, at every cross section of the guide. The natural modes in circular and rectangular guides are well known, and it has been shown[1,2] that for a cylindrical guide whose uniform cross section deviates but slightly from circular, the natural modes are only slightly different in field configuration from the corresponding modes for a circular cylinder. If however the size or shape of the cross section varies slightly from point to point, so that the guide is not strictly cylindrical, then although one may still represent the field at any given section by an expansion in terms of the natural modes of a circular cylinder, these modes are no longer independent but coupled to each other by the distortion. Electromagnetic energy in any mode will, on passing through a distorted section of guide, be partially converted into other modes, of which those above cut-off can be expected to propagate through the guide with different phase velocities from the original mode and, in the case of a physical guide, with different attenuations due to the finite conductivity of the walls. Thus the original mode exhibits an increase in effective attenuation owing to mode conversions in the distorted section, provided of course that no other mode is present at a sufficiently high level to return an appreciable amount of energy by reconversion.

Mode conversion losses become of practical importance when one attempts to transmit a mode which has intrinsically low copper loss, such as the lowest order circular electric (TE_{01}) wave in a round pipe, through a guide so large that many other modes are freely propagated at the chosen frequency. It is easy to imagine that in transmission of the TE_{01} mode through commercial copper pipe of diameter several times the wave-length, mode conversion losses might under some conditions be an appreciable fraction of the total loss; and it seems desirable to have a theoretical estimate of the magnitude of this effect. In the present paper we treat the following problem:

A pure TE_{01} wave propagating in the forward direction through a perfectly conducting circular pipe passes through a distorted section, the only restrictions on the form of the distortion being that the deviation from perfect circularity is small compared to the pipe radius, and there are no discontinuities in the surface of the pipe. No real generality is lost by the assumption of infinite conductivity, since to first approximation the losses from the TE_{01} mode due to finite conductivity and due to mode conversions are additive.* The distorted section will in general couple an infinite number of other modes to the circular electric mode; but if this section is of finite length then at sufficient distances on each side of it we shall find only those modes which propagate freely in the original circular pipe, and the power carried in each direction by these propagating modes can be calculated. The power which has been abstracted from the incident TE_{01} mode can be expressed in terms of an effective attenuation or, if desired, in terms of an average attenuation per unit length caused by mode conversions in the distorted section. In the present analysis the mode conversions are supposed to take place either in an infinitely long pipe or in one where each undesired mode is terminated

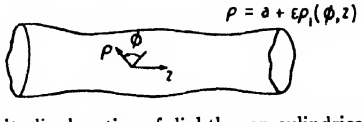


FIG. 1. Longitudinal section of slightly non-cylindrical wave guide.

[1] S. A. Schelkunoff, J. App. Phys. 9, 484 (1938).
[2] M. Jouguet, Comptes Rendus 266, 1436 (1948).

* Inhomogeneity in the conductivity of the inner wall of the guide would couple the TE_{01} mode to other modes even if the walls were perfectly smooth and circular; but an order-of-magnitude calculation has shown that for such variations in conductivity as are likely to occur in practice the coupling would be far smaller than the anticipated geometrical mode coupling effects.

whose deviation from perfect circularity is given by:

$$\epsilon\rho_1(\varphi,z)=\epsilon a f(z)\cos 2\varphi, \tag{14}$$

$$f(z)=\begin{cases}1, & 0\le|z|\le\frac{1}{2}s;\\ (l-2|z|)/(l-s), & \frac{1}{2}s\le|z|\le\frac{1}{2}l;\\ 0, & |z|\ge\frac{1}{2}l.\end{cases} \tag{15}$$

These equations represent a section of pipe, of length l, which is held with circular pipe clamps at each end and whose center section, of length s, is squeezed into an approximately elliptical shape of the same perimeter as the original pipe, the transition between the clamped ends and the distorted center section being linear. Measurements of the section of pipe used in the experiments showed that this transition was in fact very nearly linear. Of course a real pipe would not have the discontinuities in slope exhibited by (15) at $z=\pm\frac{1}{2}s$ and $z=\pm\frac{1}{2}l$; but in view of the other idealizations present in our formulation of the problem we do not feel that much would be gained by choosing a more complicated shape function with continuous slope.

In order to obtain the mode conversion losses in a pipe whose deformation is given by (14), we have to calculate the coefficients $C_{nm}^{\pm}$ and $S_{nm}^{\pm}$ defined by (8) of the preceding section. It is easily seen that all of these quantities vanish except

$$C_{2m}^{\pm}=\frac{1}{2\pi a}\int_0^{2\pi}\int_{-\frac{1}{2}l}^{\frac{1}{2}l} f(z')\exp(-i(h_{01}\mp h_{2m})z')\cos^2 2\varphi'\,dz'\,d\varphi'$$

$$=\frac{1}{a}\int_0^{\frac{1}{2}l} f(z')\cos(h_{01}\mp h_{2m})z'\,dz'$$

$$=\frac{2[\cos\frac{1}{2}(h_{01}\mp h_{2m})s-\cos\frac{1}{2}(h_{01}\mp h_{2m})l]}{(h_{01}\mp h_{2m})^2 a(l-s)}.$$

As might have been anticipated, an elliptical deformation couples only TE_{2m} modes to the TE_{01} wave. The attenuation of the circular electric mode by a single passage through the distorted section is, using Eqs. (9) to (12) of Section II,

$$N=4.343\epsilon^2\sum_m D_{2m}[|C_{2m}^{+}|^2+|C_{2m}^{-}|^2]\ \text{db}, \tag{16}$$

where the summation includes all TE_{2m} modes which are above cut-off in the circular pipe.

Using the numerical values which prevailed in the experimental work, namely

$$\lambda=3.20\ \text{cm},\quad a=2.366\ \text{in.}=6.010\ \text{cm},$$
$$l=77\ \text{cm},\quad s=17\ \text{cm},$$

we find that the only TE_{2m} waves which are above cut-off are TE_{21}, TE_{22}, and TE_{23}. A calculation the details of which will be omitted gives for the value of the sum on the right side of (16):

$$\sum_{m=1}^{3} D_{2m}[|C_{2m}^{+}|^2+|C_{2m}^{-}|^2]=38.12,$$

of which the TE_{21}^{+} mode contributes over 99.9 percent and the TE_{22}^{+} about 0.09 percent, while the TE_{23}^{+} and the three "backward" waves contribute entirely negligible amounts. Thus from (16) the attenuation per one-way trip through the elliptical section is

$$N=165.6\epsilon^2\ \text{db}. \tag{17}$$

If the distortion of the pipe is given by (14) and (15), then the parameter ϵ is related to the major and minor axes $d_{\max}$ and $d_{\min}$ of the center section by

$$\epsilon=\frac{d_{\max}-d_{\min}}{d_{\max}+d_{\min}}. \tag{18}$$

When ϵ has been determined* by (18), the theoretical attenuation is easily calculated from (17). The comparison between the values so calculated and representative experimental values, for a total of 30 round trips through the distorted section, is set forth in Table I. We regard the agreement between theory and experiment as quite satisfactory.

IV. RANDOM DISTORTIONS IN CIRCULAR GUIDE

A complete specification of the inner surface of a given section of pipe by means of the function $\epsilon\rho_1(\varphi,z)$ would allow us, at least in principle, to calculate by the formulas of Section II the mode conversion losses in that particular pipe. To obtain such detailed knowledge of $\epsilon\rho_1(\varphi,z)$ by actual measurement would, however, be highly impracticable, and at any rate we are not likely to be so interested in the properties of a single section of pipe as in the average properties of a line composed of many sections of pipe whose deviations from circularity will be more or less random. In the following paragraphs we shall calculate, for sections of pipe having a certain degree of randomness in shape, the average mode conversion loss per unit length in terms of the root-mean-square departure of the pipe from a perfect circular cylinder.

We shall begin by describing the deformation of a specific section of pipe in terms of a double Fourier series in the circumferential variable φ and the axial distance z. The deformation is thus characterized by its Fourier coefficients, which for an arbitrary continuous distortion will be a doubly infinite set of constants. In the subsequent numerical calculations, however, it will be necessary to deal with only a finite number of Fourier components; this will essentially restrict us to sections of pipe which exhibit only gradual changes in shape and no very closely spaced "ripples".

We shall proceed to express the mode conversion losses associated with the given pipe in terms of the Fourier coefficients of the distortion. We then consider

* The sample of pipe squeezed by Beck and King evidently had a radius $a=\frac{1}{4}(d_{\max}+d_{\min})$ smaller by about 0.005 in. than the nominal radius, namely 2.366 in., of the wave guide line from which it was taken. Our results should not be appreciably affected by the fact that we used the latter value of a in the calculations leading to (17).

a large number of different sections of pipe (all of equal length for simplicity), whose shapes are defined by different sets of Fourier amplitudes. We suppose that any particular Fourier coefficient is as likely to be negative as positive, so that the mean value of each coefficient, averaged over all the sections of pipe, is zero. It is also assumed that the various coefficients describing an individual section of pipe are uncorrelated, i.e., that the mean value of the product of any two different coefficients, averaged over all sections, is zero. These assumptions enable us to express the average attenuation per unit length caused by mode conversions in a wave guide line composed of many such sections in terms of the r.m.s. values of the Fourier coefficients used to describe the departure of the individual sections from perfect circularity. By making specific assumptions about the r.m.s. values of the amplitudes of the various harmonic components of distortion, we can calculate numerically the theoretical mode conversion losses in a long line.

Let us therefore consider a section of pipe whose deformation is given by a finite Fourier series:

$$\epsilon\rho_1(\varphi,z)=\tfrac{1}{4}\epsilon\sum_{s=0}^{S}\sum_{t=0}^{T}\epsilon_s\epsilon_t\left[\left(a_{st}\cos\frac{2\pi tz}{l}+b_{st}\sin\frac{2\pi tz}{l}\right)\cos s\varphi+\left(c_{st}\cos\frac{2\pi tz}{l}+d_{st}\sin\frac{2\pi tz}{l}\right)\sin s\varphi\right], \quad (19)$$

for $-\frac{1}{2}l\leq z\leq\frac{1}{2}l$ and $0\leq\varphi\leq 2\pi$. The coefficients a_{st}, b_{st}, c_{st}, and d_{st} are related to $\rho_1(\varphi,z)$ by the following scheme:

$$\begin{bmatrix}a_{st} & b_{st}\\ c_{st} & d_{st}\end{bmatrix}=\frac{2}{\pi l}\int_0^{2\pi}\int_{-\frac{1}{2}l}^{\frac{1}{2}l}\rho_1(\varphi,z)\begin{bmatrix}\cos\frac{2\pi tz}{l}\cos s\varphi & \sin\frac{2\pi tz}{l}\cos s\varphi\\ \cos\frac{2\pi tz}{l}\sin s\varphi & \sin\frac{2\pi tz}{l}\sin s\varphi\end{bmatrix}dzd\varphi. \quad (20)$$

The mean-square value of $\epsilon\rho_1(\varphi,z)$ is

$$\sigma^2=\frac{1}{2\pi l}\int_0^{2\pi}\int_{-\frac{1}{2}l}^{\frac{1}{2}l}\epsilon^2\rho_1^2(\varphi,z)dzd\varphi=\frac{\epsilon^2}{16}\left[\sum_{s=0}^{S}\sum_{t=0}^{T}\epsilon_s\epsilon_t a_{st}^2+2\sum_{s=0}^{S}\sum_{t=1}^{T}\epsilon_s b_{st}^2+2\sum_{s=1}^{S}\sum_{t=0}^{T}\epsilon_t c_{st}^2+4\sum_{s=1}^{S}\sum_{t=1}^{T}d_{st}^2\right]. \quad (21)$$

The quantities $C_{nm}^{\pm}$ and $S_{nm}^{\pm}$ which measure the mode conversion losses are given by (8) to be

$$C_{nm}^{\pm}=\frac{1}{4a^2}\sum_{t=0}^{T}\epsilon_t[a_{nt}f_{nmt}^{\pm}-ib_{nt}g_{nmt}^{\pm}],$$

$$S_{nm}^{\pm}=\frac{1}{4a^2}\sum_{t=0}^{T}\epsilon_t[c_{nt}f_{nmt}^{\pm}-id_{nt}g_{nmt}^{\pm}],$$

where

$$f_{nmt}^{\pm}=\int_{-\frac{1}{2}l}^{\frac{1}{2}l}\cos(h_{01}\mp h_{nm})z'\cos(2\pi tz'/l)dz'=\frac{(-)^t 2(h_{01}\mp h_{nm})l^2\sin\frac{1}{2}(h_{01}\mp h_{nm})l}{(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2},$$

$$g_{nmt}^{\pm}=\int_{-\frac{1}{2}l}^{\frac{1}{2}l}\sin(h_{01}\mp h_{nm})z'\sin(2\pi tz'/l)dz'=\frac{(-)^t 4\pi tl\sin\frac{1}{2}(h_{01}\mp h_{nm})l}{(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2}.$$

Hence

$$|C_{nm}^{\pm}|^2=\frac{1}{16a^4}\sum_{r=0}^{T}\sum_{t=0}^{T}\epsilon_r\epsilon_t[a_{nr}f_{nmr}^{\pm}-ib_{nr}g_{nmr}^{\pm}]\times[a_{nt}f_{nmt}^{\pm}+ib_{nt}g_{nmt}^{\pm}], \quad (22)$$

$$|S_{nm}^{\pm}|^2=\frac{1}{16a^4}\sum_{r=0}^{T}\sum_{t=0}^{T}\epsilon_r\epsilon_t[c_{nr}f_{nmr}^{\pm}-id_{nr}g_{nmr}^{\pm}]\times[c_{nt}f_{nmt}^{\pm}+id_{nt}g_{nmt}^{\pm}]. \quad (23)$$

On inserting these values into Eq. (9) and summing over n and m as in Eq. (11) we have an explicit expression for the total power converted into other modes, as a function of the Fourier coefficients a_{st}, b_{st}, c_{st}, and d_{st} describing the inner surface of the pipe.

Now let us imagine that we have a large number of sections of pipe, each of which will for the present be assumed to have the same length l. Let the kth section be represented by a finite Fourier series of the form (19), with coefficients

$$a_{st}^{(k)},\ b_{st}^{(k)},\ c_{st}^{(k)},\ d_{st}^{(k)}\quad(k=1,2,3,\cdots). \quad (24)$$

For all values of k, s, and t we may define, in accord with (20),

$$b_{s0}^{(k)}\equiv c_{0t}^{(k)}\equiv d_{s0}^{(k)}\equiv d_{0t}^{(k)}\equiv 0. \quad (25)$$

We assume that the coefficients (24) which define the various sections of pipe (excluding those which vanish identically by (25)) are *statistically independent random variables in k with mean zero and standard deviation* σ_{st}. In other words, using angular brackets to denote averages over k, so that

$$\langle f\rangle=\lim_{K\to\infty}(1/K)\sum_{k=1}^{K}f(k),$$

we assume that

$$\langle a\rangle_{st}=\langle b\rangle_{st}=\langle c\rangle_{st}=\langle d\rangle_{st}=0$$

and

$$\langle a_{st}^2\rangle=\langle b_{st}^2\rangle=\langle c_{st}^2\rangle=\langle d_{st}^2\rangle=\sigma_{st}^2, \quad (26)$$

while the average value of the product of any two coefficients with different subscripts is zero. We emphasize that these assumptions concerning the statistical behavior of the coefficients (24), while they may seem to possess a considerable degree of plausibility, are introduced at this stage as mathematical hypotheses;

TABLE II.

s \ t	0	1	σ_{st}/σ_0 2	3	4
0	25	20	15	10	5
1	20	16	12	8	4
2	15	12	9	6	3
3	10	8	6	4	2
4	5	4	3	2	1

we are not here concerned with proving their validity for any particular wave guide line.

It is easily seen from (21) and (26) that the average value of σ^2 taken over a large number of sections of pipe is given in terms of the quantities σ_{st}^2 by

$$\langle\sigma^2\rangle=\epsilon^2\left[\frac{1}{16}\sigma_{00}^2+\frac{1}{4}\sum_{t=1}^{S}\sigma_{s0}^2+\frac{1}{4}\sum_{s=1}^{T}\sigma_{0t}^2+\sum_{s=1}^{S}\sum_{t=1}^{T}\sigma_{st}^2\right], \quad (27)$$

provided that $\sigma_{st}=0$ for $s>S$ and/or $t>T$. In principle, of course, we may take S and T as large as desired.

If we add together Eqs. (22) and (23) and average the result over k, then using (25) and (26) and the assumed statistical independence of the different coefficients we obtain:

$$\langle|C_{nm}^{\pm}|^2+|S_{nm}^{\pm}|^2\rangle$$

$$=\frac{\epsilon_n}{16a^4}\left\{(f_{nm0}^{\pm})^2\sigma_{n0}^2+4\sum_{t=1}^{T}[(f_{nmt}^{\pm})^2+(g_{nmt}^{\pm})^2]\sigma_{nt}^2\right\}$$

$$=\frac{\epsilon_n l^2\sin^2\tfrac{1}{2}(h_{01}\mp h_{nm})l}{4a^4}\left\{\frac{\sigma_{n0}^2}{(h_{01}\mp h_{nm})^2l^2}\right.$$

$$\left.+4\sum_{t=1}^{T}\left[\frac{(h_{01}\mp h_{nm})^2l^2+4\pi^2t^2}{[(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2]^2}\right]\sigma_{nt}^2\right\}. \quad (28)$$

Hence finally we get from (9), (11), and (13) the following expression for the average attenuation per unit length in a line composed of a great number of sections of pipe of length l, each deviating in a random way from a perfect circular cylinder:

$$\langle\alpha\rangle=\frac{4.343\epsilon^2}{l}\sum_{n=0}^{S}\sum_{m}D_{nm}[\langle|C_{nm}^{+}|^2+|S_{nm}^{+}|^2\rangle$$

$$+\langle|C_{nm}^{-}|^2+|S_{nm}^{-}|^2\rangle] \text{ db/unit length.} \quad (29)$$

In (29) the summation over m is taken to include all the modes which are above cut-off for a given value of n (except that the TE_{01}^{+} mode is omitted).

In view of the algebraic complexity of the final expression for $\langle\alpha\rangle$, it is worth while to point out some dimensional considerations which will permit us to introduce scale factors. If we fix our attention on the attenuation per wave-length $\langle\alpha\rangle\lambda$, it is not difficult to show that (29) may be written in the form

$$\langle\alpha\rangle\lambda=F(\lambda/a,\ l/\lambda;\ \sigma_{00}/a,\ \sigma_{10}/a,\ \sigma_{01}/a,\ \sigma_{20}/a,\ \cdots), \quad (30)$$

where the function F depends only on the indicated dimensionless ratios. In other words, if we multiply all linear dimensions of the wave guide line by a given scale factor and simultaneously multiply the frequency by the reciprocal of the same factor, the attenuation per wave-length due to mode conversion losses remains constant. We shall make use of this relationship presently.

The question may be raised here whether the function F which appears in (30) depends critically upon the assumed length (in wave-lengths) l/λ of the individual sections of pipe. Physically it seems very unlikely that the mode conversion losses in an extended line whose individual sections are hundreds of wave-lengths long would be much altered by moderate fractional changes in the lengths of these sections. On the other hand, it is easy to set up idealized configurations in which the theoretical mode conversion losses do depend markedly on phase relations determined by the length of the section in which the conversions take place, and it is perhaps not *a priori* clear by inspection of (29) that similar effects do not occur in this more complicated case. To settle the question it is possible to average the right side of (29) over a selected range of values of l. Thus we find that the average attenuation per unit length owing to mode conversions in a line composed of a great number of sections of pipe whose deviations from circularity are random in the sense defined above, and whose lengths are distributed uniformly between l_1 and l_2, is

$$\langle\alpha(l_1,l_2)\rangle=\frac{4.343\epsilon^2}{4a^2(l_2-l_1)}\sum_{n=0}^{S}\sum_{m}\epsilon_nD_{nm}$$

$$\times\left[\sum_{t=0}^{T}\epsilon_t^2(I_{nmt}^{+}+I_{nmt}^{-})\sigma_{nt}^2\right] \text{ db/unit length,} \quad (31)$$

where

$$I_{nmt}^{\pm}=\frac{1}{a^2}\int_{l_1}^{l_2}l\sin^2\tfrac{1}{2}(h_{01}\mp h_{nm})l\frac{(h_{01}\mp h_{nm})^2l^2+4\pi^2t^2}{[(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2]^2}dl$$

$$=\frac{1}{4(h_{01}\mp h_{nm})^2a^2}\left\{\ln\tfrac{1}{4}|(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2|\right.$$

$$-\mathrm{Ci}[|(h_{01}\mp h_{nm})l-2\pi t|]-\mathrm{Ci}[|(h_{01}\mp h_{nm})l+2\pi t|]$$

$$+2\pi t\,\mathrm{Si}[(h_{01}\mp h_{nm})l-2\pi t]-2\pi t\,\mathrm{Si}[(h_{01}\mp h_{nm})l+2\pi t]$$

$$\left.-\frac{4\pi^2t^2\sin^2\tfrac{1}{2}(h_{01}\mp h_{nm})l}{\tfrac{1}{4}[(h_{01}\mp h_{nm})^2l^2-4\pi^2t^2]}\right\}_{l_1}^{l_2}. \quad (32)$$

Before we can calculate numerical values of $\langle\alpha\rangle$, it is necessary to reduce considerably the number of parameters involved in the general expression (29). If we knew the relative magnitudes of the mean-square Fourier amplitudes σ_{nt}^2 occurring in (28), it would be a simple matter to express their absolute magnitudes in terms of $\langle\sigma^2\rangle$ by (27) and thus to express the mode con-

version losses in terms of the latter quantity. Unfortunately information about the individual σ_{st}'s is not at present available. It may be argued that in view of the structural rigidity of commercial copper pipe the amplitudes of the higher harmonics in the distortion must tend rapidly to zero, but this tells us little about the behavior of σ_{st} for the first few terms of the series. In the absence of more detailed information and for the purpose of obtaining a numerical result, we shall arbitrarily assume that only the constant term and the first four harmonics in both the circumferential and the axial directions occur in the expression (19) for the distortion of a typical section of pipe; and we shall further assume that the r.m.s. amplitudes of these harmonics are given by

$$\sigma_{st}=(5-s)(5-t)\sigma_0,\ 0\leq s,\ t\leq 4, \tag{33}$$

where σ_0 is a length proportional to the r.m.s. deviation of the wave guide line from an average circular cylinder. The relative magnitudes of the non-vanishing σ_{st}'s are thus assumed to be as shown in Table II:

It is easy to determine from (27) and (33) that the mean-square deviation of the wave guide line from an average cylinder is given by

$$\langle\sigma^2\rangle=1314.1\epsilon^2\sigma_0^2.$$

Numerical calculations have been carried out, under the assumptions just described, to obtain the mode conversion losses in transmission of TE_{01} power through a long wave guide composed of 20-foot sections of 5-inch (outside diameter) pipe at both X-band and K-band wave-lengths. For the X-band case we take

$$\lambda=3.20\text{ cm},\quad \lambda/a=0.532,$$
$$a=2.366\text{ in.},\quad l/\lambda=190.5.$$
$$l=20\text{ ft.},$$

Substitution of these values into (29) gives, after a lengthy calculation,

$$\langle\alpha\rangle=4.92\langle\sigma^2\rangle/a^2\text{ db/cm}, \tag{34}$$

where $\langle\sigma^2\rangle$ and a^2 are expressed in the same units. The numerical coefficient in (34) represents the combined effect of 14 forward modes (excluding TE_{01}^+) and 15 backward modes; but it turns out that the TE_{31}^+ and TE_{21}^+ modes together account for 97.6 percent of the total (TE_{31}^+, 85.8 percent; TE_{21}^+, 11.8 percent). It will in general be true that pipe deformations which are long compared to the wave-length convert power from the TE_{01} mode mostly into those (transverse electric) modes whose phase velocities are nearest the TE_{01}. Physically it is evident that unless the phase velocities of two modes are nearly equal, they will phase in and out so rapidly that over a distance of several wave-lengths the net power transfer will be small.*

If the r.m.s. deviation $\langle\sigma^2\rangle^{\frac{1}{2}}$ of the pipe from a perfect circular cylinder is expressed in mils, then since $a=2.366$ in., (34) becomes

$$\langle\alpha\rangle=0.879\times10^{-6}\langle\sigma^2_{\text{mils}}\rangle\text{ db/cm}. \tag{35}$$

This value may be compared with the theoretical attenuation α_0 due solely to heat losses. We find from the usual wave guide formulas[8] that for the TE_{01} mode at 3.2 cm in a copper pipe of diameter 4.732 in. and resistivity 1.72×10^{-6} ohm cm,[9]

$$\alpha_0=1.076\times10^{-5}\text{ db/cm}, \tag{36}$$

so that the ratio of mode conversion losses to theoretical heat losses, each expressed in db/cm, is

$$\langle\alpha\rangle/\alpha_0=0.0817\langle\sigma^2_{\text{mils}}\rangle. \tag{37}$$

A discrepancy of 8 percent of the theoretical copper losses could be accounted for by mode conversions, under the assumptions that led to (34), if the r.m.s. deviation of the pipe from perfect circularity were 1 mil. We may point out that if the circumferential variations in the pipe surface were sinusoidal, a 1-mil value of $\langle\sigma^2\rangle^{\frac{1}{2}}$ as here defined would correspond to a maximum variation of 2.83 mils in the radius, or, in the case of an even harmonic, to a difference of 5.66 mils between maximum and minimum internal diameters.

We should stress that the numbers just obtained depend rather strongly upon the assumptions which we made concerning the relative magnitudes of the various σ_{st}'s; in particular, the results are sentitive to variations in the relative amounts of second and third harmonics in the circumferential distortion, since these harmonics couple the two modes, TE_{21} and TE_{31}, whose phase velocities are closest to TE_{01}. It may be of interest to record the results of a calculation made under the assumption that all of the σ_{st}'s corresponding to harmonics of degree not greater than four are present in equal amounts; i.e.,

$$\sigma_{st}\begin{cases}=\sigma_1,\ 0\leq s,\ t\leq 4,\\ =0,\text{ otherwise.}\end{cases} \tag{38}$$

Then we find from (27) that

$$\langle\sigma^2\rangle=18.06\epsilon^2\sigma_1^2,$$

and from (29) and (36), after extensive computation,

$$\langle\alpha\rangle/\alpha_0=0.205\langle\sigma^2_{\text{mils}}\rangle,$$

which is two and one-half times the value of $\langle\alpha\rangle/\alpha_0$ given by (37). Although the "tapered" distribution of amplitudes (33) seems physically much more reason-

* On the other hand, unpublished calculations have indicated that the effect of a surface irregularity which is *short* compared to the wave-length is to couple the TE_{01} mode most strongly to those modes which are near cut-off. With such an irregularity one might expect pronounced resonance phenomena if the frequency were varied through the cut-off frequency of a particular higher mode.

[8] Reference 4, p. 390.

[9] The effective resistivity of metallic surfaces at centimeter wave-lengths is appreciably higher than the d.c. resistivity of the bulk metal, because of surface roughness and corrosion. See E. Maxwell, J. App. Phys. 18, 629 (1947), and S. P. Morgan, Jr., J. App. Phys. 20, 352 (1949). In estimating mode conversion losses from measurements of total line attenuation it is first necessary to take account of these augmented heat losses.

TABLE III. Values of the coefficients D_{nm} for $\lambda\ a=0.532$.

m \ n	0	1	2	3	4
			D_{nm}		
1	1.7305	1.085	3.770	8.592	16.27
2	6.824	7.363	13.38	22.97	38.21
3	22.77	23.86	43.16	112.3	—
4	—	244.3	—	—	—

TABLE IV. Values of the coefficients D_{nm} for $\lambda/a=0.208$.

m \ n	0	1	2	3	4
			D_{nm}		
1	0.2401	0.1562	0.5325	1.798	2.146
2	0.8208	0.9710	1.642	2.514	3.607
3	1.7830	2.499	3.559	4.845	6.384
4	3.208	4.858	6.400	8.230	10.39
5	5.249	8.270	10.47	13.07	16.19
6	8.207	13.16	16.35	20.24	25.06
7	12.78	20.41	25.43	31.97	40.98
8	21.16	32.36	41.96	—	—

able than the "flat" distribution (38), it must be remembered, in view of our lack of experimental measurements, that the actual distribution in commercial copper pipe may be quite different from either of these assumed distributions.

In order to compare mode conversion losses in the same pipe at different frequencies, we turn now to K-band wave-lengths and take

$$\lambda=1.25\text{ cm},\quad \lambda/a=0.208,$$
$$a=2.366\text{ in.},\quad l/\lambda=488.$$
$$l=20\text{ ft.},$$

If we assume, as in obtaining (34), that the σ_{si}'s are given by (33), a calculation similar to the preceding one gives

$$\langle\alpha\rangle=2.94\langle\sigma^2\rangle/a^2\text{ db/cm},\qquad (39)$$

or, if $\langle\sigma^2\rangle^{\frac{1}{2}}$ is expressed in mils,

$$\langle\alpha\rangle=0.525\times10^{-6}\langle\sigma^2{}_{\text{mils}}\rangle\text{ db/cm}.\qquad (40)$$

In deriving the numerical coefficient of Eq. (39) contributions have been included from the 37 forward modes whose phase velocities are nearest to that of the TE_{01} mode; the contributions of the other modes would be too small to affect the third significant figure. It appears during the calculation that 98.2 percent of the mode converted power is carried by five modes (TE_{31}^+, 51.8 percent; TE_{21}^+, 27.3 percent; TE_{12}^+, 12.4 percent; TE_{11}^+, 4.7 percent; TE_{11}^+, 2.0 percent).

Equations (39) and (40) may be compared respectively with (34) and (35); we see that in absolute magnitude the mode conversion losses at 1.25 cm are only 0.597 times the losses due to mode conversions in the same pipe at 3.2 cm. But the copper losses for the TE_{01} mode diminish even more rapidly as the frequency is increased;* at 1.25 cm the theoretical attenuation of this mode in a 4.732-inch pipe is

$$\alpha_0=2.51\times10^{-6}\text{ db/cm}.$$

Hence the ratio of mode conversion loss to theoretical heat loss at 1.25 cm is

$$\langle\alpha\rangle/\alpha_0=0.209\langle\sigma^2{}_{\text{mils}}\rangle,\qquad (41)$$

and this is 2.56 times the relative mode conversion loss at 3.2 cm.

As a check on the preceding calculations, we have determined from (31) and (32) the average mode conversion power carried by the TE_{21}^+ and TE_{31}^+ modes at 3.2 cm for pipe sections ranging in length from 15 to 25 feet. Thus substituting

$$\lambda=3.20\text{ cm},\quad l_1=15\text{ ft.},$$
$$a=2.366\text{ in.},\quad l_2=25\text{ ft.},$$

we find that

$$\langle\alpha(l_1,l_2)\rangle\approx4.67\langle\sigma^2\rangle/a^2\text{ db/cm},\qquad (42)$$

where the right side is a very close lower limit to the actual value of $\langle\alpha(l_1,l_2)\rangle$. Similarly, taking

$$\lambda=1.25\text{ cm},\quad l_1=15\text{ ft.},$$
$$a=2.366\text{ in.},\quad l_2=25\text{ ft.},$$

and considering the five modes nearest in phase velocity to TE_{01}, we find

$$\langle\alpha(l_1,l_2)\rangle\approx3.03\langle\sigma^2\rangle/a^2\text{ db/cm}.\qquad (43)$$

On comparing (42) with (34) and (43) with (39) we see that the results obtained by considering 20-foot sections of pipe are within a few percent of the average results for lengths between 15 and 25 feet.

Finally we may inquire the effect of mode conversions on the TE_{01} wave at K-band in a smaller guide, say a 2-inch pipe instead of 5-inch. We get approximately this case if we scale down the results for X-band in a 5-inch pipe, using the dimensional relation (30) and the calculations which led to (34). The latter equation may be written

$$\langle\alpha\rangle\lambda=15.75\langle\sigma^2\rangle/a^2\text{ db/wave-length},\qquad (44)$$

and this relation holds for the values

$$\lambda=1.25\text{ cm},\quad a=0.924\text{ in.},\quad l=7.81\text{ ft.},$$

since the ratios λ/a and l/λ are the same as those for which it was derived. But at $\lambda=1.25$ cm, (44) gives

$$\langle\alpha\rangle=12.60\langle\sigma^2\rangle/a^2\text{ db/cm},$$

or, if $\langle\sigma^2\rangle^{\frac{1}{2}}$ is expressed in mils,

$$\langle\alpha\rangle=14.76\times10^{-6}\langle\sigma^2{}_{\text{mils}}\rangle\text{ db/cm}.$$

Since the theoretical copper losses for the TE_{01} mode at 1.25 cm in a guide of inside diameter 1.848 in. lead to an attenuation

$$\alpha_0=4.41\times10^{-5}\text{ db/cm},$$

we find for the ratio of mode conversion losses to heat losses,

$$\langle\alpha\rangle/\alpha_0=0.335\langle\sigma^2{}_{\text{mils}}\rangle.\qquad (45)$$

* In practice however surface roughness effects become more significant as the skin depth for eddy currents is reduced.

Comparing (45) with (41), we see that when we go from a 4.732-inch pipe to a 1.848-inch pipe at 1.25 cm, we increase the relative mode conversion losses, for a fixed value of $\langle\sigma^2_{\text{mils}}\rangle$, by a factor of 1.60.

ACKNOWLEDGMENTS

I wish to thank Mr. A. C. Beck and Mr. A. P. King for permission to quote certain of their unpublished experimental results. I am indebted to Dr. W. A. Tyrrell for helpful suggestions during the course of this work, and to Mrs. Jean A. Smith for carrying out most of the numerical calculations.

APPENDIX I

Excitation of Transverse Electric Modes by Tangential Electric Field at Boundary

Let the tangential components of the electric field vanish over the surface $\rho=a$ of a circular wave guide everywhere except in a finite region S where they are given by

$$E_\varphi=E(a,\varphi,z), \quad E_z=0. \tag{A1}$$

The stream function $\Psi(\rho,\varphi,z)$ corresponding to these assigned boundary values may be found by the following formal procedure:

Let the stream function giving the boundary value

$$E_\varphi=E_0(a,\varphi,z)=\delta(\varphi-\varphi')\delta(z-z'), \tag{A2}$$

where δ is the unit impulse-function, be

$$\Psi_0(\rho,\varphi,z;\varphi',z')$$
$$=\sum_{n=0}^{\infty}\left[\int_{(C)} A_n(\gamma)J_n[(\beta^2+\gamma^2)^{\frac{1}{2}}\rho]e^{\gamma z}d\gamma\right]\frac{\epsilon_n}{2\pi}\cos n(\varphi-\varphi'),$$

where the contour (C) in the γ plane will be specified presently and $A_n(\gamma)$ is a function to be determined. Now

$$E(a,\varphi,z)=[\partial\Psi_0/\partial\rho]_{\rho=a}=\sum_{n=0}^{\infty}\left[\int_{(C)} A_n(\gamma)(\beta^2+\gamma^2)^{\frac{1}{2}}J_n'[(\beta^2+\gamma^2)^{\frac{1}{2}}a]e^{\gamma z}d\gamma\right]\frac{\epsilon_n}{2\pi}\cos n(\varphi-\varphi'). \tag{A3}$$

But if we write $\delta(\varphi-\varphi')$ as a Fourier series in φ and $\delta(z-z')$ as a contour integral,[10] (A2) becomes

$$E_0(a,\varphi,z)=\lim_{\substack{s\to 0\\ t\to 0}}\sum_{n=0}^{\infty}\left[\frac{1}{2\pi i}\int_{(C)}\frac{\sinh\frac{1}{2}\gamma s}{\frac{1}{2}\gamma s}\right.$$
$$\left.\times\exp[\gamma(z-z')]d\gamma\right]\frac{\epsilon_n}{2\pi}\frac{\sin\frac{1}{2}nt}{\frac{1}{2}nt}\cos n(\varphi-\varphi'), \tag{A4}$$

the contour (C) in the γ-plane being along the imaginary axis indented to the right at the origin, as in Fig. 3. We determine $A_n(\gamma)$ by comparing (A3) and (A4), and obtain, on letting s and t pass to zero under the signs of summation and integration, the result

$$\Psi_0(\rho,\varphi,z;\varphi',z')$$
$$=\sum_{n=0}^{\infty}\left[\frac{1}{2\pi i}\int_{(C)}\frac{J_n[(\beta^2+\gamma^2)^{\frac{1}{2}}\rho]\exp[\gamma(z-z')]d\gamma}{(\beta^2+\gamma^2)^{\frac{1}{2}}J_n'[(\beta^2+\gamma^2)^{\frac{1}{2}}a]}\right]\frac{\epsilon_n}{2\pi}$$
$$\times\cos n(\varphi-\varphi')$$
$$=\sum_{n=0}^{\infty}\frac{\epsilon_n I_n}{2\pi}\cos n(\varphi-\varphi'),$$

where

$$I_n=\frac{1}{2\pi i}\int_{(C)}\frac{J_n[(\beta^2+\gamma^2)^{\frac{1}{2}}\rho]\exp[\gamma(z-z')]d\gamma}{(\beta^2+\gamma^2)^{\frac{1}{2}}J_n'[(\beta^2+\gamma^2)^{\frac{1}{2}}a]}. \tag{A5}$$

It is easily shown that the integrand in (A5) is a single-valued function of γ, so that I_n can be evaluated in terms of the residues at the poles of the integrand. These poles are at the points* $\gamma=\pm\Gamma_{nm}$, where $k_{nm}=(\beta^2+\Gamma_{nm}^2)^{\frac{1}{2}}a$ is the mth positive root of $J_n'(x)=0$ and Γ_{nm} is the corresponding natural propagation constant in the guide. The contour (C) is indented to the right of the poles on the negative imaginary γ-axis and to the left of the poles on the positive imaginary γ-axis; then if $z>z'$ it is closed with an infinite semi-circle in the left half-plane, thus including an infinite number of poles on the negative real axis. With the aid of l'Hospital's rule and Bessel's equation it is easy to show that

$$\lim_{\gamma\to-\Gamma_{nm}}\frac{\gamma+\Gamma_{nm}}{J_n'[(\beta^2+\gamma^2)^{\frac{1}{2}}a]}=\frac{k_{nm}^3}{\Gamma_{nm}a^2(k_{nm}^2-n^2)J_n(k_{nm})},$$

so that the contribution of the pole at $\gamma=-\Gamma_{mn}$ to the stream function $\Psi_0(\rho,\varphi,z;\varphi',z')$ is

$$\frac{k_{nm}^2\epsilon_n J_n(\chi_{nm}\rho)\cos n(\varphi-\varphi')\exp[-\Gamma_{nm}(z-z')]}{2\pi\Gamma_{nm}a(k_{nm}^2-n^2)J_n(k_{nm})} \tag{A6}$$

for $z>z'$. If $z<z'$ the contour (C) is to be closed with an infinite semicircle in the right half-plane; the end result

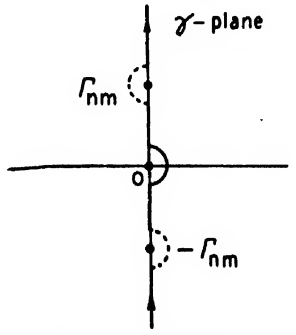


FIG. 3. The contour (C) in the γ-plane.

[10] Reference 4, pp. 34–35.

* For $n=0$ there are also poles at $\gamma=\pm i\beta$, but the contribution of these poles to the stream function is of the form constant $\times\exp(\pm i\beta z)$, which leads to no non-vanishing fields.

is to change the sign of the exponent in the last factor of (A6). Hence in general

$$\Psi_0(\rho,\varphi,z;\varphi',z')=\sum_{n=0}^{\infty}\sum_{m=1}^{\infty}\frac{\epsilon_n k_{nm}^2 J_n(\chi_{nm}\rho)\cos n(\varphi-\varphi')\exp[-\Gamma_{nm}|z-z'|]}{2\pi\Gamma_{nm}a(k_{nm}^2-n^2)J_n(k_{nm})}. \tag{A7}$$

Having the stream function which corresponds to an impulse-function boundary value of E_φ, we are able to write down at once the stream function for the general field given by (A1); it is

$$\Psi(\rho,\varphi,z)=\int\int_S F(a,\varphi',z')\Psi_0(\rho,\varphi,z;\varphi',z')dz'd\varphi',$$

which corresponds by (A7) to Eq. (5) of Section II.

APPENDIX II

Some values of the quantities D_{nm} defined by (10) have been calculated for particular choices of λ/a and are given in Tables III and IV.

Wettability, a Function of the Polarizability of the Surface Ions

L. R. Sonders, D. P. Enright, and W. A. Weyl
The Pennsylvania State College, School of Mineral Industries, State College, Pennsylvania
(Received October 28, 1949)

The wettability of crystals, glasses, and even of water itself can be temporarily decreased by bringing ions of high polarizability into their surfaces. Base exchange experiments are described where the hydrogen ions present in the surface layers of bentonite (a clay mineral of high exchange capacity) and of a soda-lime glass are replaced by different cations. This substitution seems to have no particular effect on the hydrophilic and rheological properties of the carriers as long as their surfaces are kept in contact with water. After drying, however, the surfaces which contain ions of high polarizability become hydrophobic, at least temporarily. A porous clay film will no longer absorb water instantaneously after having been treated with Ni^{2+}, Mn^{2+}, Hg^{2+}, or similar ions with incomplete outer electron shells. Contact angles with water up to 70° could be observed for a short period for Hg^{2+} and Pb^{2+} clays.

Glass capillary tubes which have been treated with non-noble gas-type ions show a capillary rise which is much smaller than that observed with the tubes which have been treated with HCl and water only. This depression of the capillary rise, too, is temporary and can be observed only if the glass wall has been thoroughly dried previous to the experiment and if the capillary rise is measured in the upward direction.

An explanation is presented for these and allied phenomena on the basis of the polarization of ions in the strongly asymmetrical forcefields of interfaces. The experiments are correlated with the hysteresis of the contact angle and with observations concerning adhesion phenomena and catalytic activities of heavy metal ions at interfaces.

I. INTRODUCTION

THE present difficulty in the understanding and in the correct interpretation of surface phenomena, such as heterogeneous catalysis, adhesion and wetting properties of solids, is partly the result of not being able to directly measure the surface forces emanating from a solid without changing the surface structure. For example, the forcefields emanating from the surfaces of alkali halides cannot be measured by bringing the salts onto a platinum surface and then determining their adhesion, because a metal surface in contact with NaCl is not identical with one in contact with KI. It is now well established that the electrical properties of metal films are changed if they are brought in contact with different gases or solids. The checkered forcefield of the cube face of NaCl or of a glass surface polarizes the metal surface by inducing higher electron densities opposite the cations and lower electron densities opposite the anions. The metals respond to the external forcefield by producing image forces and these in turn affect the surface structures of the alkali halides.

This uncertainty about the forcefields of surfaces and their change with the environment is not generally appreciated. A surface of galena (PbS), for example, exerts different forces upon its environment depending on its chemical composition. The galena surface undergoes a structural change if one medium, say air, is replaced by another, for example, water. The forcefields of the highly polarizable Pb^{2+} and S^{2-} ions can assume directional properties. This structural change in the galena surface affects its contact angle with a liquid and, as a result, the contact angle depends on whether the liquid is advancing over or receding from the solid surface. Adam[1] in his authoritative book writes on this hysteresis phenomenon as follows: ". . . mathematical works on Capillarity rarely mention it at all. It may, however, be observed without any apparatus with almost any rain drop traveling down an ordinarily dirty window pane. This difference between advancing and receding angles is often called the 'hysteresis' of

[1] N. K. Adam, *The Physics and Chemistry of Surfaces* (Oxford University Press, London, 1941), third edition, p. 180.

the contact angle. Early references to it in the literature are scarce; Rayleigh referred to it in 1890; Pockels measured it for several liquids; Sulman studied it in connection with the flotation of minerals. It is well known to those who have measured surface tension by the rise in a capillary tube, for the rise is often less if the liquid is pushing its way up from below than if it is falling from a point above its final equilibrium height.

"The cause of the 'hysteresis' of contact angles is still obscure. It looks superficially as if there was a frictional resistance to motion of the liquid edge over the solid."

It is the object of this paper to explain this hysteresis of the contact angle. Experiments with solid-liquid and liquid-gas interfaces are described and their explanation is based on the polarizability of ions and the effect which the asymmetrical forcefield of an interface exerts upon the direction of valency forces.

A note on this subject has been published as a Letter to the Editor.[1a]

II. THE POLARIZABILITY OF IONS AND SOME OF ITS MANIFESTATIONS

The polarizability of a molecule is frequently measured by its response to the alternating electromagnetic field of light. It can be easily computed from refractive index, molecular weight, and density. The polarizability is a measure for the change which takes place in the outer electronic orbits if an atom, molecule, or ion is exposed to an external electrical field. The polarizability of an ion depends on its size, charge, and on its electron configuration. The rules governing polarization of ions have been derived by Fajans[2] and his school.

Without going into details, some of these rules shall be mentioned.

(a) Anions are more polarizable than cations of the same configuration. For example, the ions which have the electron structure of the inert gas atom, neon, show polarizabilities which decrease with increasing nuclear charge.

$$O^{2-} > F^{-} > Ne > Na^{+} > Mg^{2+}.$$

(b) For ions of the same electrical charge and of comparable electron configuration, the polarizability increases with the size (Table I).

(c) The polarizability of cations of approximately equal size is larger for those of 18 outer electrons than for those of the noble gas type, but is greatest for those cations which have only one or two outer electrons.

Sr^{2+}(8 outer electrons) $< Cd^{2+}$(18 outer electrons) $< Pb^{2+}$(18+2 outer electrons).

The mercurous ion Hg^{+} has only one outer electron and its polarizability is, therefore, extremely high.

The polarizability of an ion can be described as its ability to adjust its own forcefield to that of the environment. This concept has been used by Weyl[3] for explaining adhesion phenomena and solubilities involving two incompatible media. A metal, for example, ordinarily is not soluble in a fused glass. Glasses, say soda-lime-silicates, consist only of noble gas-type ions having rather low polarizabilities. If a strongly polarizable cation such as Pb^{2+}, Bi^{3+}, Sn^{2+}, or Tl^{+} is introduced into the glass, a metal such as gold or copper becomes soluble to a certain extent. These ions act as "metallophilic groups" comparable with hydrotropic substances or with emulsifiers. In the asymmetrical forcefield, which exists between the neutral metal atom and the negatively charged O^{2-} ion, metallophilic ions are strongly polarized. Their electron density no longer possesses the spherical symmetry of the ion in the gaseous state, but has become strongly asymmetrical as far as the outer or valency electrons are concerned.

The deformation of a strongly polarizable ion, such as a Pb^{2+} ion, can be described in the language of the chemist as a disproportionation. Its inner shells (*K* to *O* shell) are not directly affected by chemical changes. It is the *P* shell of a lead atom with its four 6*s* electrons which is primarily responsible for its chemical properties. The four outer electrons of the *P* shell can be donated to electron acceptors. The result of this electron transfer is the formation of the fourfold positively charged Pb^{4+} ion. If only two of the four valence electrons are surrendered to an electron acceptor a Pb^{2+} ion is formed. Outside of the 18 electrons of the *O* shell, such a Pb^{2+} ion, has still two 6*s* electrons. On approach of a negatively charged oxygen ion these two outer electrons are repelled. As a result, the electrical or valence forces of the Pb^{2+} ion are no longer equally distributed in space. The one side of the Pb^{2+} from which the two electrons are repelled has a lower electron density and resembles in its structure the Pb^{4+} ion (18 outer electrons). The

TABLE I. Molar refraction of gaseous ions in cc (after Bauer and Fajans).[a]

		He	Li^{+}	Be^{2+}
		0.521	0.08	—
O^{2-}	F^{-}	Ne	Na^{+}	Mg^{2+}
6.95	2.44	1.00	0.47	0.26
S^{2-}	Cl^{-}	A	K^{+}	Ca^{2+}
22.7	9.06	4.20	2.25	1.40
Se^{2-}	Br^{-}	Kr	Rb^{+}	Sr^{2+}
28.8	12.66	6.39	3.79	2.58
Te^{2-}	I^{-}	Xe	Cs^{+}	Ba^{2+}
40.9	19.21	10.43	6.53	4.73

[a] See reference 2(a).

[1a] Enright, Sonders, and Weyl, J. App. Phys. 20, 1011 (1949).

[2] (a) K. Fajans, *Chemical Forces and Optical Properties of Substances* (McGraw-Hill Book Company, Inc., New York, 1931). (b) N. Bauer and K. Fajans, J. Am. Chem. Soc. 64, 3023 (1942).

[3] W. A. Weyl, J. Soc. Glass Tech. 32, 247–259 (1948).

TABLE II.

Glass type	Gain in weight (1 cm² glass powder in 4 hr.)
Lead borate (51.7 percent PbO)	4.9 mg
F. 620 (45 percent PbO)	5.6 mg
Window glass	32.0 mg
Bottle glass	35.0 mg

TABLE III. Influence of adsorbed cations on the capillary rise of water in glass tubes.[a]

Adsorbed cation	Example for the capillary rise in mm for different treatments				Average decrease in rise (percent) for different treatments		
	Blank	24 hr.	48 hr.	120 hr.	24 hr.	48 hr.	120 hr.
Ba^{2+}	62	62	59	56	0	5.4	9.2
Tl^{+}	63	63	63	57	0	0	10.2
Fe^{2+}	60	56	55	53	8.5	10.8	12.5
Co^{2+}	64	60	57	55	5.8	10.5	13.1
Zn^{2+}	61	57	55	53	6.9	11.7	13.8
Cu^{2+}	64	58	55	55	9.0	13.3	14.4
Cd^{2+}	60	57	52	50	5.4	13.5	15.7
Cr^{3+}	63	58	53	53	7.5	13.9	10.1
Mn^{2+}	62	55	52	50	10.1	16.5	18.0
Ni^{2+}	62	55	49	48	12.2	20.2	21.8
Hg^{2+}	60	46	40	40	24.1	32.8	32.8
Pb^{2+}	61	44	39	37	29.0	38.1	40.8

[a] The value for the blank (H^+ glass) varied between 60 and 64 mm depending on the diameter of the tube. All data presented are averages of several experiments carried out with different tubes.

other side characterized by a higher electron density resembles the neutral Pb atom.

One can describe this deformation or polarization of the lead ion by the scheme:

$$Pb^{2+} = \tfrac{1}{2}Pb^{4+} + \tfrac{1}{2}Pb.$$

This is analogous to other "internal oxidation-reduction" reactions such as:

$$2Sn^{2+} = Sn^{4+} + Sn \quad \text{or} \quad Hg_2Cl_2 = HgCl_2 + Hg.$$

The strongly polarized Pb^{2+} ion, therefore, exerts strong ionic forces toward the oxygen ion at its one side but is "neutral" at the opposite side. The side to which the two 6*s* electrons have been repelled has the property of atomic lead and, as a result of its high electron density at this side, it may form a "metallic bond" with a metal atom and, thus, keep atomic Au or Cu in solution.

A similar state of polarization was found to exist for Pb^{2+} and other strongly polarizable cations in surface layers. The polarization of surface ions affects the surface tension, hygroscopicity, catalytic properties, as well as adhesion forces.

Summarizing, we may say that polarizable ions change their electron distribution and with it their "chemical affinities" when brought into the asymmetrical forcefield of a surface or of an interface. In contrast to the gaseous ion or to the hydrated ion which has an electron density function of spherical symmetry, the polarized surface ion exerts directional valence forces because its electron density distribution has become asymmetrical.

III. WETTABILITY OF A SOLID SURFACE

Whether or not a solid brought into a liquid goes into solution depends on its lattice energy and the heat of solvation. Solubility is an equilibrium phenomenon and, as such, it is fairly well understood. There seems to be no relationship, however, between the solubility of a crystal and its wettability by the solvent. Anhydrous chromium chloride may be cited as a typical example. This compound is soluble in water and forms several hydrates, when allowed to crystallize from its aqueous solutions. Nevertheless, it is not easily wet by water and one cannot obtain a reproducible contact angle between $CrCl_3$, H_2O, and air. Even the uncontaminated freshly sublimed anhydrous $CrCl_3$ will float on water for hours without being wet. This phenomenon is due to the state of polarization of the large Cl^- ions in the surface layer where they are exposed to the positive forcefields of Cr^{3+} ions on one side only. The Coulomb forces of these polarized surface Cl^- ions are directed only toward the interior and, therefore, they are not available for hydration. The air film covering the $CrCl_3$ is gradually replaced by water, and the presence of "contrapolarizing" cations is required in order to make the $CrCl_3$ crystal hydrophilic. Addition of HCl accelerates this process because now the Cl^- ions come into a more symmetrical forcefield; they have a Cr^{3+} ion on the one side and a H^+ ion on the other.

Just as striking is the behavior of freshly precipitated mercuric iodide HgI_2. This compound contains the two very polarizable ions Hg^{2+} and I^-. When precipitated from aqueous solution the iodide behaves like any other hydrophilic precipitate, for example, $BaSO_4$, AgCl, etc. However, on filtering, the HgI_2 changes into a hydrophobic material as soon as it comes in contact with air. In the asymmetrical forcefield crystal-air, the polarizable ions adjust their own forcefields to that of the interface by directing their valence forces primarily toward the interior of the crystal. This structural change causes the surface-free energy of the crystal to be lowered.

The low surface-free energy or the lack of strong attraction forces of glasses containing some strongly polarizable heavy metal ions causes them to be less hygroscopic than others which contain only noble gas-type ions of low polarizability. This phenomenon was discovered by Hubbard.[4]

When developing a more logical test for optical glasses which correspond to actual service conditions instead of treating a lens with hot nitric acid, Hubbard found that the hygroscopicity of otherwise "chemically poor" glasses can be much lower than that of a bottle glass of known high chemical resistivity as can be seen from Table II. Glasses containing only noble gas-type ions exert strong surface forces because their low polarizability makes it impossible for them to adjust their Coulomb forces to the asymmetrical forcefield of the

[4] D. Hubbard, J. Research Nat. Bur. Stand. 36, 365–375 (1946).

surface. A surface ion of such a glass has incomplete coordination and extends unsaturated valence forces into space. As a result, oncoming water molecules have a greater chance to be attracted and chemisorbed at such a glass surface: the glass is hygroscopic. However, a glass containing Pb^{2-} ions does not extend strong forces toward the atmosphere and, as a result, only a smaller number of collisions with oncoming H_2O molecules will lead to adsorption.

The difference between a lead glass and a soda-lime glass is not very convincing evidence that it is the polarizability of the surface ions which determines primarily the wettability of a solid and, as such, is responsible for the hysteresis of the contact angle. Experiments were carried out, therefore, with a soda-lime silicate glass and a clay of high base exchange capacity, both of which can be easily changed in respect to surface forces by bringing different cations into their surfaces.

IV. CAPILLARY RISE IN GLASS TUBES CONTAINING DIFFERENT CATIONS AT THEIR INNER SURFACE

The surface of a soft (soda-lime-silicate) glass in contact with water consists of OH^- groups. These hydroxy-groups extend hydrogen bonds toward any liquid containing one or more OH^- groups (water, glycerol, sulfuric acid) so that these liquids completely wet the glass and give a zero contact angle. The importance of the OH^- groups for the wettability of a glass surface can be easily demonstrated.

Concentrated sulfuric acid to which a trace of potassium dichromate has been added is heated in a test tube. The chromate oxidizes organic impurities and produces a clean glass surface which is completely wetted by the liquid. Sulfuric acid molecules can attach themselves to this surface by hydrogen bonds. Adding a trace of sodium fluoride to the sulfuric acid causes the OH^- groups to be replaced by F^- ions according to the equation:

$$-\overset{|}{\underset{|}{Si}}-OH+HF=-\overset{|}{\underset{|}{Si}}-F+H_2O.$$

The dehydrating effect of the concentrated sulfuric acid causes this reaction to proceed to completion. After a few seconds, the inner surface of the test tube has become depleted of OH^- groups and is now covered with fluorine ions. With the OH^- groups missing, the surface lacks the ability to form hydrogen bonds with the H_2SO_4 molecules and, as a result, the sulfuric acid forms little droplets on the glass surface and behaves like water in a greasy container. The substitution of F^- for OH^- has completely changed the interfacial forces between glass and sulfuric acid, because the fluorine ion under the strongly polarizing influence of the silicon ion cannot exert a Coulomb force upon the H_2SO_4 molecules which is sufficiently strong to overcome the cohesion forces of the liquid and to produce wetting.

TABLE IV. Temporary hydrophobicity imparted to a clay film by adsorbed ions.

Cation	Approximate contact angle	Average time for spreading
H^+	0°	Immediate spreading
Li^+	0°	Immediate spreading
Na^+	0°	Immediate spreading
K^+	0°	Immediate spreading
Mg^{2+}	20°	1–3 min.
Ca^{2+}	20°	1–3 min.
Sr^{2+}	20°	1–3 min.
Be^{2+}	20°	1–3 min.
Al^{3+}	20°	1–3 min.
Ba^{2+}	40°	5–10 min.
Tl^+	40°	5–10 min.
Zn^{2+}	40°	5–10 min.
Cd^{2+}	40°	5–10 min.
Cu^{2+}	40°	5–10 min.
Co^{2+}	40°	5–10 min.
Cr^{3+}	40°	5–10 min.
Ni^{2+}	60°	10–12 min.
Mn^{2+}	60°	10–12 min.
Pb^{2+}	70°	15–30 min.
Hg^{2+}	70°	15–30 min.
Laurylamine	80°	No spreading after 120 min.

In the absence of water the silicon fluoride layer is stable and, if the glass tube with the sulfuric acid is sealed, the inner surface retains its repellency indefinitely. However, in the presence of water, hydrolysis takes place and the OH^- groups are restored. The glass surface again is completely wet by water; that means it has a zero contact angle.

In contact with the aqueous solution of a salt some of the H^+ ions of the glass surface can be replaced by other cations; for example, Hg^{2+} or Pb^{2+}. This base exchange is accomplished by treating glass capillaries of about 0.3-mm inner diameter first with diluted HCl and, after rinsing them with water, exposing them to a salt solution for several days. After rinsing with distilled water and drying at 110°C, the glass surface was modified by adsorbed cations. The number of ions adsorbed on the glass depends on the time of exposure, the nature of the compounds, its concentration, and on the temperature. Rinsing of the glass after the adsorption is necessary in order to remove the excess salt and the loosely adsorbed anions. Unfortunately, this procedure introduces a certain error because the substitution of H^+ by other cations is a reversible reaction.

The capillary rise was then determined by carefully inserting the treated glass tubes into water and measuring the rise in the *upward direction.*

The salts containing cations of the noble gas-type did not exert a measurable effect on the capillary rise, namely, the nitrates of Li^+, Na^+, K^+, Be^{2+}, Mg^{2+}, Ca^{2+}, Sr^{2+} and Al^{3+}. The effect of other cations can be seen from Table III.

The experiments on the capillary rise of water in glass tubes which had been treated with various cations

shows how the polarizability of the cation determines the temporary interfacial tension value. With the exception of the large and, therefore, polarizable Ba^{2+} ion, all cations with eight outer electrons (noble gas-type ions) did not affect the wettability of the surface to a noticeable extent. As the capillaries were allowed to stand in the water a gradual rise was observed and the time required for reaching the equilibrium value varied from one ion to another. It was greatest for Hg^{2+} (144 hr.) and Pb^{2+} (120 hr.), less for Ni^{2+} (72 hr.) and Cd^{2+} (48 hr.).

As water molecules distill from the meniscus onto the glass surface, the forcefield of the polarized surface ions gradually becomes more symmetrical and with increasing symmetry the strength of attraction forces emanating from the surface ions increases. As a result the liquid rises, replaces the gas film as well as the adsorbed cations and gradually the original glass surface (OH^- groups) is restored.

V. WETTABILITY OF CLAY SURFACES

The same phenomenon of "temporary hydrophobicity" can be observed for clay minerals. Because of its high base exchange capacity, Wyoming bentonite was selected for these experiments. The clay was purified by electrodialysis and allowed to stand with the electrolyte for 48 hr. under occasional agitation by a mechanical stirrer. The concentration of the electrolyte was chosen so as to provide 200 milliequivalents of the cation per 100 grams of the dry clay.

The homogeneous clay suspensions containing the various cations adsorbed at their surface were then sprayed onto porous ceramic tiles by means of a DeVilbiss spray gun. This method produced smooth clay films of controlled thickness. The porosity of the tile removes both the excess electrolyte and water from the clay film.

The clay-covered tiles were then dried at 110°C for 24 hr. Their wettability was tested by dropping freshly distilled water upon the clay, estimating the contact angle of the water droplet right after its application and determining the time necessary for its complete disappearance.

As the bentonite shows practically no anion exchange capacity it was immaterial which salt was used for introducing the cations. The experiments showed that no significant difference exists between nitrates, chlorides, and sulfates. Nitrates were chosen for the base exchange experiments.

Table IV summarizes the results. The cations are divided into several groups according to their increasing effectiveness based on the contact angle. Table IV brings out the parallelism which exists between the effect of these ions on the wettability of a clay film and on the capillary rise in glass tubes.

In order to show the effect of a long chain hydrocarbon group on the wettability and the spreading time, a clay was treated with a solution of the organic base, laurylamine. As far as the contact angle is concerned, Pb^{2+} and Hg^{2+} ions come very close to this substance. However, in contrast to the temporary hydrophobicity imparted by the polarized cation, the lauryl group causes the clay surface to become permanently water repellent.*

VI. WETTABILITY OF A WATER SURFACE

The phenomenon that a drop of a liquid, for example, one falling from a reflux condenser onto the surface of a boiling liquid, does not immediately unite, but may be seen floating for several seconds, is well known to every chemist and has been described in the textbooks on surface chemistry. It can be explained on the basis that gas films or vapor films are present at the surface. The removal of adsorbed gas molecules or the rearrangement of adsorbed vapor molecules so that they fit into the structure of the liquid are time consuming processes.

We have found that polarizable cations, such as Pb^{2+} or Hg^{2+} dissolved in water affect its surface properties in a manner similar to surface active agents, such as sodium stearate or laurylamine. It can be demonstrated that there is no basic difference between the effects of mercuric acetate and sodium stearate upon the forcefield emanating from a water surface. However, the long chains of the stearic acid molecules cannot turn around very rapidly, but the state of polarization of the Hg^{2+} ions change instantaneously with the symmetry of the forcefield. As a result, the similarity between the effects of the two compounds can be demonstrated best by a dynamic rather than by a static method.

The influence of polarizable cations on the surface forces of water affects the tendency of a stream of water flowing out of an orifice to trap air bubbles. Distilled and degassed water was allowed to flow from a burette into a glass container. Both burette and container were equipped with overflow tubes which ensured constancy of the two water levels. By a screw attachment the container could be gradually lowered or raised and the variable distance between the orifice of the burette and the water level could be measured. As long as the tip of the burette was kept close to the water level of the container a steady flow could be observed and no gas bubbles were trapped. By gradually increasing the distance between the tip of the burette and the water level of the container, one reaches a point where gas bubbles become entrapped and carried under the water surface. It was found that for one and the same apparatus and orifice this critical height, H, could be reproduced with fair accuracy. Taking the essential precautions, such as using boiled and degassed water, avoiding contamination and using degreased glass ware (no stopcock grease), the values for H were reproducible within one percent.

* Viscosity measurements of the different clay suspensions did not reveal any unusual features for those clays which contained Pb^{2+} or Hg^{2+} at their surfaces.

The phenomenon can be explained as follows. Depending on the height and on the speed of the flowing water a surface film is formed which has its own characteristic structure. Its structure is the result of two antagonistic effects, namely, the tendency of the system to assume a molecular arrangement which decreases its surface-free energy and the thermal motion favoring a random orientation.

Gas molecules are being adsorbed at the water surface and their desorption is a time consuming process. The desorption of the gas film, however, is essential before the surface of the streaming water can unite with that of the stagnant water. For slowly flowing water the rate of combination of the two water surfaces is slower than the gas desorption. However, as H and with it the time of exposure of the water to air and the velocity of the falling water is increased, a point is reached where the speed of the uniting water surfaces exceeds the rate of desorption. As a result the gas film is carried below the water level where it assumes spherical shape under the influence of the surface tension: gas bubbles are formed.

The addition of sodium stearate to the flowing or to the stagnant water or to both causes their surface structures to change, the surface tension of the water is lowered. In other words, the ratio of attraction forces toward air to attraction forces toward the liquid is increased if a surface active compound such as sodium stearate is added to the water. A one-tenth molal sodium stearate solution was found to trap the airfilm at a height which was nearly 20 percent lower than that for the pure water. A solution of laurylamine (two percent by volume) decreased the value of H nearly 15 percent.

This method of studying the surface forcefields of solutions was found to be very sensitive as can be seen from Table V which shows the effects of a 0.5 molal and a 1.0 molal salt solution on the critical height, i.e., the one above which a gas film is trapped and carried under the water surface. In contrast to the experiments on the wettability of glass and of clay, it was found that even the Na^+ and K^+ ions showed some influence on the wettability of water. In addition, it was found that the nature of the anion is not without influence as can be seen from a comparison between the values for a sodium nitrate and a sodium acetate solution.

A 1.0 molar solution of $NaNO_3$ causes H to decrease 1.2 percent due to the weak polarizability of the hydrated sodium ion. The NO_3^- ion has no effect as can be seen from the addition of HNO_3. The sodium acetate in 1.0 molal solution was found to decrease the value of H by 2.4 percent which means that the acetate group must have about the same effect as the Na^+ ion.

The response of this method to changes in the constitution of aqueous solutions makes it an interesting tool for studying complex formation and related phenomena. For example: depression of H for 0.25 molar mercuric chloride, 9.1 percent; depression of H for 0.25 molar potassium chloride, 0; depression of H for 0.25 molar KCl+0.25 m $HgCl_2$, 4.8 percent.

TABLE V. Effect of dissolved salts on trapping of airfilm by flowing water.

Cation	Salt Anion	0.5 molal solution H mm	0.5 molal solution Percent decrease	1.0 molal solution H mm	1.0 molal solution Percent decrease
H^+	NO_3^-	83	0	83	0
Li^+	NO_3^-	83	0	83	0
Na^+	NO_3^-	83	0	82	1.2
K^+	NO_3^-	82	1.2	79.5	4.2
Be^{2+}	NO_3^-	83	0	83	0
Mg^{2+}	NO_3^-	83	0	83	0
Ca^{2+}	NO_3^-	82.5	0.6	80.5	3.0
Sr^{2+}	NO_3^-	81.5	1.8	79	4.8
Ba^{2+}	Br^-	80	3.6	75	9.7
Al^{3+}	NO_3^-	83	0	83	0
Zn^{2+}	NO_3^-	81	2.4	78	6.0
Cu^{2+}	NO_3^-	80.5	3.0	77	7.2
Cd^{2+}	NO_3^-	79	4.7	74.5	10.2
Fe^{3+}	NO_3^-	79.5	4.2	74	10.9
Mn^{2+}	NO_3^-	78	6.0	71	14.5
Cr^{3+}	NO_3^-	78	6.0	70.5	15.1
Co^{2+}	NO_3^-	77	7.2	70.5	15.1
Ni^{2+}	NO_3^-	76.5	7.7	69.5	16.3
Pb^{2+}	NO_3^-	70.5	15.1	54.4	34.4
Tl^+	CH_3COO^-	69.5	16.3	52	37.3
Hg^{2+}	CH_3COO^-	65.5	21.1	50	39.8

These data indicate that the highly polarizable Hg^{2+} ion which is responsible for the strong decrease of H is made less effective in the presence of KCl. The chemistry of these compounds suggests the formation of $(HgCl_4)^{2-}$ ions or of a similar complex anion which is less polarizable than the cation Hg^{2+}.

For experiments of this type, it is important to control the temperature fairly accurately because an equilibrium such as

$$HgCl_2+2KCl \rightleftharpoons K_2(HgCl_4)$$

is temperature sensitive. In general it was found that the temperature is not a major factor. For our experiments, no special precautions had to be taken to control the temperature accurately because even a 10°C deviation from room temperature would have caused H to vary less than one percent. For pure water, H was found to increase 2.4 percent if the temperature was lowered from 32° to 3°C.

VII. CONCLUSIONS

All surface phenomena such as adhesion, wetting of solids by liquids, and heterogeneous catalysis are governed by the polarization properties of ions. Perfect crystals, liquids, or gases provide a symmetrical environment for ions and molecules at least as far as their bulk is concerned. Interfaces, such as solid-gas or liquid-gas, are the seat of highly asymmetrical forcefields which cause the polarizable ions to undergo a strong change in their electronic orbits. This change will not look very spectacular when described by the elec-

tron density function of the atom because it affects only the tail end of this curve, i.e., the part which represents the outer orbitals. However, these orbitals are the seat of optical and chemical properties and, consequently, they determine light absorption, fluorescence, electrical conductivity, as well as oxidation and reduction phenomena. An ion which is exposed to the one-sided polarizing influence of an interface becomes optically and chemically a different individual. It has different light absorption and different fluorescence. If the polarizability of an ion is sufficiently high it may lose those properties which are the result of its electrical excess charge. A Tl^{+} or Pb^{2+} ion is easily hydrated, because their excess charges orient and attract the dipoles of water. However, the same ions adsorbed at a dry clay or a dry glass surface have an electron density distribution which is different from the ions in their gaseous state or from the ions when symmetrically surrounded by water. The O^{2-} ions of the carrier repel the valency electrons of these cations thus increasing their forcefield toward the solid and decreasing their forces emanating into space. The polarized cations lower the surface free energy of the solid and by becoming neutral towards the exterior—the repelled electrons counteract the effect of the positive excess charge —they have lost the property to attract water molecules. It is this decrease of the Coulomb forces directed toward the exterior which causes the solid to appear water repellent. Strictly, there are no repulsion forces acting in the surface which make the carrier hydrophobic but the attraction forces have become too weak to overcome the cohesion of the liquid.

Solids containing large and, therefore, polarizable ions, e.g., I^{-}, S^{2-}, etc., especially when crystallizing in layer lattices (PbI_2, CdI_2, MoS_2), are not easily wet and due to their weak forcefields they assume properties usually associated with lubricants.

In contrast to organic materials with long hydrocarbon chains, the inertness of polarized ions is only temporary. Every one who has walked along a dusty road on a dry hot summer day and been caught by a sudden rain may have observed this phenomenon of temporary hydrophobicity. The dry dust is not easily wet by water and the first rain drops may roll around like mercury. However, after it has rained a few minutes the rain soaks into the ground and soon the mud tells us that the dust did not have permanent hydrophobic properties.

Polarizable ions can adjust their own forcefields to that of the environment. Their Coulomb forces are symmetrical when symmetrically surrounded by ions of the opposite charge. In the asymmetrical forcefield of an interface, however, their electron distribution and with it their valency forces assume directional properties. A heavy metal ion adsorbed at a surface which repels its electron cloud, so that the polarized ion has a higher electron density toward the exterior, assumes the properties of a neutral metal atom. As such, it may be used as a bridge between an ionic substance, for example, glass, and the metals silver, gold, or copper. Polarizable ions are used to enhance the adhesion between glass and metal. Before silvering a glass it is treated with tin compounds in order to make the silver film adhere better.

Enright, Marboe, and Weyl[5] used the decomposition of H_2O_2 into H_2O and oxygen for demonstrating the metallic character of adsorbed heavy metal ions. Hydrogen peroxide is readily decomposed when brought in contact with a noble metal, Au or Pt. Dissolved ions, such as Hg^{2+} or Ag^{+}, have no effect on this reaction. However, if these ions are adsorbed at the surface of an ionic carrier (glass or alumina) they become positive catalysts. Ions of this type act as negative catalysts or catalytic poisons when adsorbed at the surface of platinum, because in its forcefield they form metallic bonds with the metal (high electron density between cation and metal surface) and extend a strong positive forcefield toward the solution which is ineffective as a catalyst for this reaction.

The experiments with flowing water prove that the water-gas interface has a similar deforming influence on polarizable ions at the glass or clay surface. Very probably this interface plays an important role in certain gas-liquid reactions which are catalyzed by heavy metal ions. The rate of oxidation of ferrous sulfate by air or oxygen gas is greatly increased if 0.01 molar cupric ions are present. In a similar fashion the hydration of acetylene gas in diluted acid leading to acetaldehyde is catalyzed by Hg^{2+} ions.

[5] Enright, Marboe, and Weyl, "Influence of the carrier on the catalytic activity of adsorbed heavy metal ions," Tech. Report No. 1, Contract No. N6onr 269 Task Order 8 N.R. 032-265 (October, 1948).

Distribution of Ionization in Materials Irradiated by Two and Three Million-Volt Cathode Rays

J. G. Trump, K. A. Wright, and A. M. Clarke
Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received November 4, 1949)

Measurements are reported on the distribution of ionization in depth of aluminum produced by steady beams of two and three million-volt electrons. The variation of cathode-ray current density in a plane transverse to the beam, the effect on this transverse distribution of additional aluminum scattering foils, and a practical method of cathode-ray dosage computation are given.

INTRODUCTION

In the study of the physical and biological effects of high energy cathode rays[1] it is desirable to know the actual distribution of the ionization energy in the absorber material. An earlier paper[2] reported measurements on the distribution of ionization in depth of water, aluminum, copper, and lead produced by monoenergetic electron beams over the energy range from 0.3 to 1.5 million volts. In the present studies, ionization distribution in depth of aluminum and intensity distribution across the beam were measured for steady streams of electrons in the two and three million-volt range. In order to secure greater uniformity across the field, a study was made of the effect of scatter produced by aluminum foils inserted in the beam. A method of dosage computation which facilitates the determination of maximum, average, and minimum cathode-ray dose on a sample of given size is included.

CATHODE-RAY SOURCE

A pressure-insulated Van de Graaff electrostatic accelerator[3] was used as the source of high energy electrons. These were directed toward the absorber in a continuous stream, homogeneous in energy and controllable over the range from 1 to 3.5 million volts. The electrons emerged from the tube through an aluminum window 0.003 inch thick on which they impinged as parallel particles over an area of about 5 mm diameter. Voltage was measured by a generating voltmeter which had been calibrated at 1.63 million volts by making use of the beryllium (γn) reaction and observing the threshold with a BF_3 counter.

MEASUREMENT TECHNIQUE

A thin parallel-plate ionization chamber was used to measure the distribution of ionization in depth caused by the high energy electrons in their passage through aluminum. The chamber consisted essentially of a stretched aluminum diaphragm 0.0006-inch thick, insulated and separated by 1 mm from a thick aluminum plate which formed the other electrode. The area of the air gap was large compared with that of the cathode-ray beam at the region of measurement. As shown in Fig. 1, the chamber was mounted close beneath the cathode-ray window of the accelerator tube. Thin aluminum sheets were inserted between the cathode-ray window and the ionization chamber to give the effect of placing the chamber at progressively greater depths below the surface of the absorbing material.

A constant beam current I_1 of high energy electrons was directed at the absorber system. Two microamperes of beam current produced maximum ionization currents I_2 or I_3 of approximately 40 μa. I_2 represents the negative ionization current plus the portion of the beam current which passed through the aluminum sheets, and I_3 the positive ionization current minus the portion of the beam current which was absorbed in the aluminum sheets.

Linearity of the ionization chamber was checked by using a fixed thickness of aluminum absorber above the chamber and observing the constancy of the ratio I_2/I_1

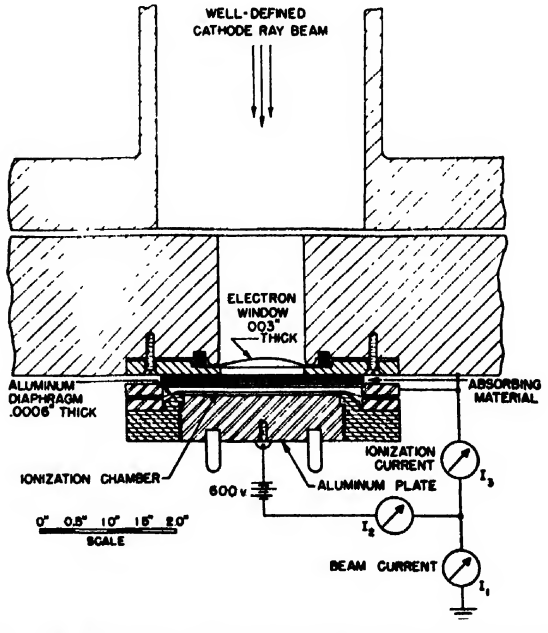


FIG. 1. Apparatus for measuring the distribution of ionization in depth of aluminum produced by high energy electron beams.

[1] J. G. Trump and R. J. Van de Graaff, J. App. Phys. 19, 599 (1948).
[2] Trump, Van de Graaff, and Cloud, Am. J. Roentgenology and Rad. Therapy 43, 728 (1940).
[3] J. G. Trump and R. W. Cloud, Am. J. Roentgenology and Rad. Therapy 49, 531 (1943).

over a wide range of beam currents I_1. The collecting voltage was maintained constant at 600 volts, this voltage having been found to be well above the amount for saturation. This check indicated linearity of response to within ± 2 percent for the range of currents used in the experiment.

The distribution of current density in a plane transverse to the beam was measured with the arrangement shown in Fig. 2 by collecting the current passing through spaced holes in an aluminum plate at a distance of 40 cm from the exit window of the accelerator tube. The minimum or inherent scattering of the electron beam was caused by both the 0.003-inch aluminum window and the 40 cm of intervening air. Faraday cages were placed beneath the individual holes spaced radially at $\frac{1}{4}$-inch intervals from the center of the plate. Holes which did not have Faraday cages beneath were plugged to eliminate stray currents. The $\frac{1}{8}$-inch diameter holes in the lower aluminum plate were drilled so

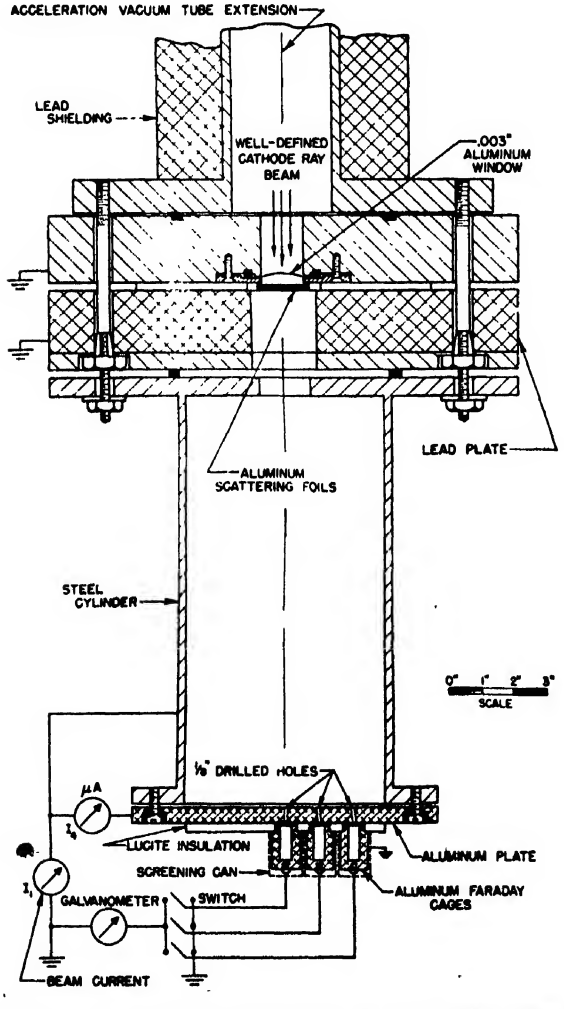


FIG. 2. Apparatus for determining the cathode-ray current density transverse to the beam and the effect of additional aluminum scatterer.

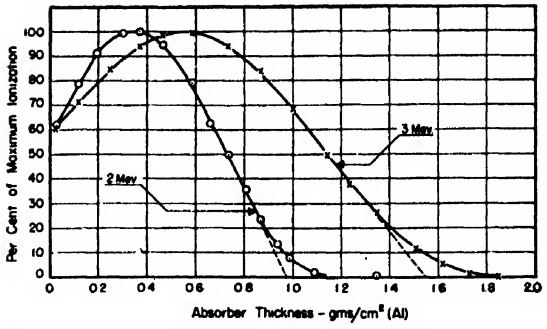


FIG. 3. Distribution of ionization in depth of aluminum produced by cathode rays with energies of two and three million volts.

that their angles with respect to the beam axis were determined by the center of the exit window and the radial distance of the holes from the center of the plate. It was found necessary to shield the individual Faraday cages from ionization in air caused by x-rays. For these measurements the total beam current I_1 was held constant at 50 μa, and the resultant Faraday cage currents of the order of 10^{-8} amp. were read on a galvanometer.

IONIZATION DISTRIBUTION IN DEPTH

Figure 3 shows the ionization distribution in depth of aluminum produced by electron beams of two and of three million volts energy. It is observed that the ionization has a broad maximum at about one-third the greatest range, this previously reported distribution[2] arising from the high scattering tendency of the primary electrons within the absorber. The penetrating power of electrons in aluminum is close to 1 g/sq. cm per two million volts, as determined by the interpolated curve intercept. From other studies[2] at lower electron energies, a corresponding range was found for water, aluminum, copper, and lead, and the general shapes of the ionization distribution curves were closely similar. The range in centimeters can be obtained by dividing the range in g/sq. cm by the density of the absorber material. The tail of the distribution curve is due primarily to those few electrons which have made relatively undeviated paths through the absorber. Ionization due to x-rays produced within the absorber can easily be shown to constitute a negligible addition to the measured ionization currents.

Since the area of the ionization chamber was large in comparison with that of the beam, the curves of Fig. 3 represent approximately the distribution of ionization in depth along the axis of the beam and more accurately the total ionization at each depth. In addition to its primary dependence on electron energy and absorber density, the distribution of ionization energy in an absorber irradiated by essentially parallel and monoenergetic electrons is profoundly affected by electron scattering. Although the ionization density is greatest near the end of the path of a single electron,

the effect of scattering on a beam of such particles is to shift the region of highest ionization and energy absorption to a depth of about one-third the maximum penetration for that energy.

BEAM INTENSITY DISTRIBUTION ACROSS THE FIELD

Figures 4 and 5 show the intensity distribution of two and three million-volt cathode-ray beams in a transverse plane 40 cm from the exit window. The upper curve of each figure shows the distribution resulting from the inherent scatter consisting of the 0.003-inch aluminum window and 40 cm of atmospheric air. The remaining curves show the beam-spreading effects of additional aluminum scatterer placed immediately below the window.

The greater directivity and reduced scatter obtained at the higher beam energies is evident from a comparison of Figs. 4 and 5. In the irradiation of an area four inches in diameter at the 40-cm distance, two million-volt

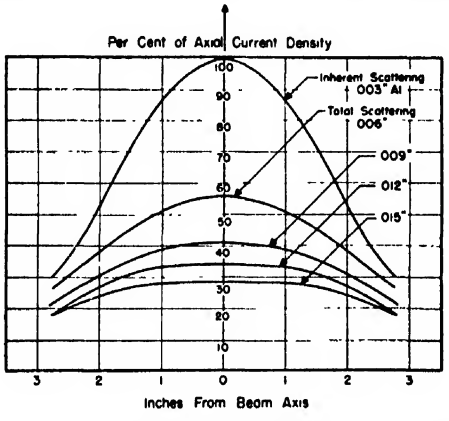


FIG. 4. Distribution of current density for two million-volt cathode rays in a transverse plane 40 cm from the 3-mil aluminum window.

cathode rays with inherent scatter would deliver 52 percent of the axial intensity to the sample edges, while a three million-volt beam would deliver but 30 percent to the edges. Additional scatterer inserted in the beam rapidly increases the uniformity of beam intensity over the selected area, but is attended by loss of charge to outside regions. Modification of the intensity distribution to the transverse plane can readily be accomplished by the addition of solid scatterer or an increase in the air path to the absorber. Because the ratio of scatter to absorption increases with atomic number, it would be preferable to use foils of gold or other high atomic-number metal as the scattering material.

DOSAGE COMPUTATIONS

At the present time there exists a considerable interest in the physical, chemical,[4] and biological[5] effects of high energy cathode rays on living and non-living materials. For such irradiation studies, the cathode-ray equipment may be set up as in Fig. 2 with a solid lower aluminum plate. A current-collecting disk is placed on, and preferably flush with, the top of this plate, but electrically insulated from it. A suitable shutter is used to provide accurate control of the exposure time.

In principle, the cathode-ray dosage rate can be specified with great exactness inasmuch as it is determined by the number of impinging electrons per second and their accelerating voltage. In practice, account must be taken of the distribution of the ionization energy both in depth and across the transverse plane of the absorber. For the case of stationary samples, dosage computations can readily be made with the aid of information contained in Figs. 3–5. The dose delivered to liquids in continuous flow or to solids on a moving belt may often be derived more simply by direct comparison with the known chemical or bactericidal effects of cathode rays as determined on stationary samples.

The total cathode-ray power falling on a disk of diameter D is EI where E is the accelerating potential and I the total beam current to the disk. The average power P absorbed per gram of material distributed uniformly over this disk to a depth R expressed in g/sq. cm is

$$P=[EI/\pi(D/2)^2R]K_1K_2\text{watts/g},$$

where K_1 is the fraction of the total power absorbed in range R and K_2 is a factor close to unity which takes account of the possible change in electron backscatter from the disk area after a sample of the same diameter has been placed upon it. The value of K_1 may be obtained from Fig. 3 by dividing the area under the curve for a sample thickness R by the total area of the curve of energy E. The value of K_2 requires the measurement of electron backscatter for different materials,[6] but can be kept close to unity by making both disk and sample

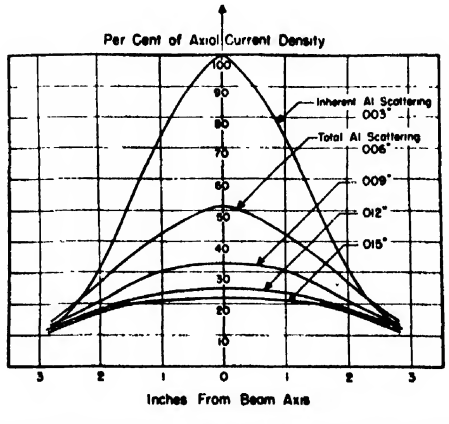


FIG. 5. Distribution of current density for three million-volt cathode rays in a transverse plane 40 cm from the 3-mil aluminum window.

[4] A. O. Allen, AEC Report MDDC-363 (September, 1946).
[5] Dunn, Campbell, Fram, and Hutchins, J. App. Phys. 19, 605 (1948).
[6] J. G. Trump and R. J. Van de Graaff, Phys. Rev. 75, 44 (1949).

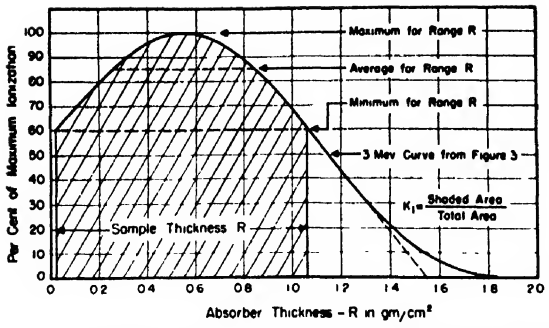


FIG. 6. Ionization-in-depth curve for three million-volt cathode rays applied to a 1.03-g/cm² sample to determine maximum, average, and minimum doses.

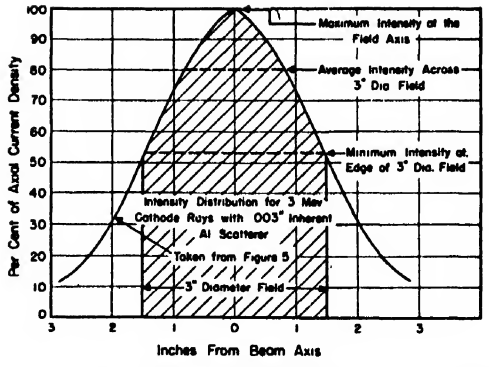


FIG. 7. Transverse beam intensity curve for three million-volt cathode-ray beam applied to a three-inch diameter sample at 40-cm distance to determine maximum, average, and minimum intensities.

cover of the same low atomic number material, such as aluminum.

Cathode-ray dose is best expressed in ergs and joules for physical studies, but may be specified in equivalent roentgens for biological studies. The conversion factors for these units of radiation energy are 1 joule $= 10^7$ ergs; 1 roentgen equivalent physical[7] (rep) $= 83$ ergs per gram of tissue; 10^6 reps $= 8.3$ joules per gram of tissue. The irradiation time is found by dividing the desired dose expressed in joules by the average dosage rate P.

Maximum and minimum doses can be calculated from the average by applying the appropriate dose ratios both for the distribution in depth and the transverse distribution as given in Figs. 3–5. Consider the specific case in which it is desired to irradiate a stationary sample three inches in diameter (7.62 cm) with three million-volt cathode rays. The current to the three-inch collecting disk is taken as 10 μa. For convenience, the sample depth R in g/sq. cm is chosen so that the entrance and exit doses are the same percent of the maximum dose. From the three million-volt curve of Fig. 3 or 6, R is found to be 1.03 g/sq. cm. The ratio of the area included under this curve to depth R to the total area gives coefficient $K_1 = 0.84$. If care has been taken to use low atomic-number materials and to mount the sample for minimum effect on backscatter, then K_2 can be taken as unity. Then

$$P = 3\times10^6 \times \frac{10\times10^{-6}}{\pi(7.62/2)^2} \frac{0.84\times1}{1.03} = 0.54 \frac{\text{watts}}{\text{g}}$$

and the time required to deliver an average dose of 10^6 reps is

$$t = 8.3/0.54 = 15.4 \text{ sec.}$$

From the distribution-in-depth curve of Fig. 6 the average dose is found to be 85 percent of the maximum, and the minimum 60 percent. From the intensity distribution across the beam of Fig. 7, with only inherent scattering, the average intensity is found to be 80 percent of the maximum, and the minimum 52 percent. Therefore the minimum dose is $(0.6\times0.52)/(0.85\times0.8)$, or 46 percent of the average, and the maximum dose is $(1\times1)/(0.85\times0.8)$, or 147 percent of the average. The spread between maximum and minimum doses may be controlled by suitable selection of absorber depth, width, scatterer, and irradiation distance.

ACKNOWLEDGMENTS

We desire to pay tribute to Arthur M. Clarke, our associate on this paper, whose death on January 19, 1949, ended seven years of devoted and effective work in the High Voltage Research Laboratory. The help of Dr. E. W. Webster in making some of the measurements is gratefully acknowledged. This work was assisted by the Joint Program of the ONR and the AEC.

[7] R. D. Evans, Nucleonics 1, 39 (October, 1947). Proposals that 1 rep be associated with the absorption of 93 ergs per gram of water or tissue and that the value of 83 ergs per gram be restricted to absorption in air will be considered at the Sixth International Congress of Radiology in London, July 1950.

Letters to the Editor

Metal Transfer in the Cutting Process*

MILTON C. SHAW AND CHARLES D. STRANG, JR.
Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts
January 27, 1950

THERE are generally thought to be two distinct types of continuous chip formed in single point metal cutting—that produced in the presence of a so-called built-up edge and that produced when no built-up edge is present. The built-up edge is a metal appendage which accumulates at the cutting point of the tool when the metal is cut under conditions conducive to high friction and welding. The built-up edge can be found adhering to either the tool face or the workpiece and is frequently macroscopically visible. Figure 1 is a photo-micrograph of a polished

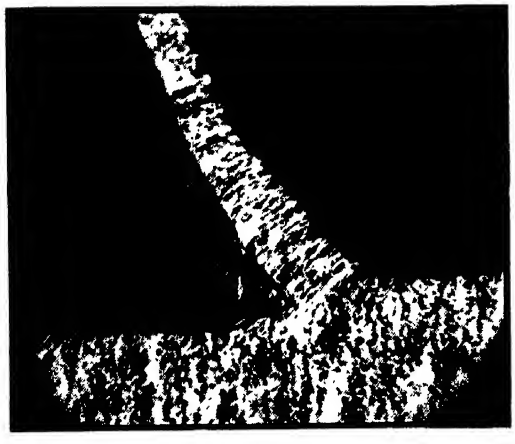


FIG. 1. Photo-micrograph showing built-up edge, magnification, 60 times.

and etched specimen obtained by cutting a sample of annealed SAE 4027 steel in air with the end milling cutter shown in Fig. 2. The large built-up edge can be seen at the junction of the chip and the freshly cut surface.

The chip shown in the photo-micrograph of Fig. 3 was produced under conditions least favorable to the formation of a built-up edge. The workpiece and the tool were submerged in carbon tetrachloride, one of the most effective cutting fluids known, and the cutting velocity was adjusted to be in the range where the carbon tetrachloride has sufficient time to react (below approximately 40 ft. per min.) with the metal being cut to form a low shear strength solid lubricant. This compound is thought to prevent or decrease metal-to-metal contact between chip and tool face and so prevent formation of a built-up edge. No built-up edge is evident in Fig. 3 nor was any transferred metal visible on the cutting tool when it was examined under the microscope. In spite of this, the autoradiograph of Fig. 4 shows that considerable material from the workpiece has been welded to the face of the tool.

This autoradiograph was produced by machining a specimen of SAE 4027 steel which had previously been made radioactive by neutron irradiation in the pile at Oak Ridge. A new tool, identical to the first, was used for this experiment and was carefully cleaned and washed in water to remove any loose chips or residual radioactive iron chloride produced as the metal was machined under carbon tetrachloride. The cutting end of the tool was placed in contact with a sheet of x-ray film and exposed for a period of two

FIG. 2. End milling cutter.

weeks. Figure 4 is a positive print and the white areas indicate the presence of the radioactive built-up edge. It will be seen that the built-up edges does not extend across the full width of the tool face inasmuch as the workpiece was annular in cross section.

This simple experiment establishes the fact that even under the most ideal conditions of cutting fluid action, some welding occurs between the chip and tool face with attendant transfer of metal and further that the presence or absence of a built-up edge is merely a matter of degree. While the invisible built-up edge indicated by Figs. 3 and 4 is insignificant from the standpoint of cutting geometry, it is of considerable interest with regard to friction and tool wear. Although carbon tetrachloride does not entirely prevent metal to metal contact and welding at the chip-tool interface, it does act to reduce greatly the amount of welding or transfer and the size of the built-up edge. The iron chloride that is formed when carbon tetrachloride reacts with the metal during cutting probably reduces the real area of metal-to-metal contact and at the same time may produce a transferred layer that is weak in shear, consisting in part of steel and iron chloride.

Under usual cutting conditions, the built-up edge is welded to the tool face whereas it is generally found attached to the chip when the operation is suddenly stopped for the purpose of preparing photo-micrographs as shown in Fig. 1. This paradox may be readily explained on the basis of decreased shear strength with increased temperature. During cutting, the temperature at that surface of the built-up edge furthest from the tool face is higher than that at the tool face and slip consequently takes place in the region of temperature-reduced shear strength. When a photo-micrograph specimen is prepared by suddenly stopping the tool in the act of producing a chip, the metal quickly cools and the shear strength of the highly worked metal at the interface of the built-up edge and the chip exceeds that at the tool face. Therefore, when the tool is withdrawn, the built-up edge is usually found adhering to the chip rather than to the tool. On the other hand, remnants of built-up edge are often found on the face of a tool which has continued cutting to the very end of the workpiece.

The relatively large amount of metal transfer (built-up edge) which occurs when cutting metal in air as compared to that observed in dry friction experiments[1] may also be explained by the

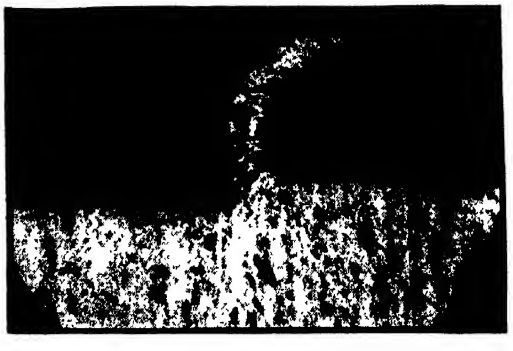


FIG. 3. Photo-micrograph of chip formed beneath the surface of carbon tetrachloride. Magnification, 60 times.

FIG. 4. Autoradiograph of end milling cutter used to machine radioactive steel.

decrease in shear strength with increase in temperature. When cutting metal, considerable energy is expended in plastic deformation along the shear plane. The energy thus dissipated causes an increase in the internal energy along the shear plane, a consequent rise in temperature and hence a decrease in shear strength. The portion of the material in the built-up edge adjacent to the tool face is at a lower temperature and hence has a greater shear strength. The presence of the plastic deformation in the shearing process thus acts to allow the transfer of an unusually large quantity of metal from the chip to the tool face before the built-up edge produced becomes unstable and passes off with the chip.

In the field of metal cutting research, as in many other branches of physics and engineering, the use of radioactive materials provides a means for detecting changes of very small magnitude. This technique should prove particularly valuable in studies of the chemical reactivity of cutting fluids, the general nature of cutting fluid action, and in detecting very small rates of tool wear.

* The radioactive steel used in these experiments was irradiated at Oak Ridge under authorization of the United States AEC and was made available through the generosity of Professor J. T. Burwell, Jr. and the Chrysler Corporation.

[1] J. T. Burwell, Jr., "Radioactive tracers in friction studies," Nucleonics 38 (December, 1947).

The Potential Distribution in Pulsed Oxide-Coated Cathodes and Its Consequences for the Velocity Distribution of the Emitted Electrons

R. LOOSJES, H. J. VINK, AND C. G. J. JANSEN
Philips Research Laboratories, N. V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands
January 23, 1950

OUR investigations[1] concerning the total potential drop across a pulsed, activated oxide* coating of 25 micron thickness showed that a potential difference of e.g. 200 v at a current density of 10 amp./cm² at 827°C can exist.

Further investigations[2] by means of probe potential-measurements showed that, when using a rather pure nickel core containing a few tenths percent of different impurities, there was no appreciable interface potential difference and only a moderate potential drop of e.g. 6 v at 10 amp./cm² at 900°C across 85 microns of an oxide coating of 270 micron total thickness was set up during a pulse.

From these results it was concluded that, at least under the given circumstances, the potential distribution in a pulsed, activated oxide coating is of the form reproduced in Fig. 1. It was supposed that the rather steep potential drop near the surface could give rise to a spread of the electron velocities over tens or hundreds of volts, (a) because of electrons starting not only from points on the surface like *A*, but also emerging from points more inside the porous coating like *B*, or (b) because of all the electrons starting from points like *A*, while the potential at these points would be rather variable in connection with local variations in the properties of the surface layer of the coating. We therefore worked out a method for investigating the electron velocity distribution by means of electrostatic deflection.[3] Great care was taken not to lose any group of electrons especially those with the smaller velocities.

Preliminary results showed that the spread of the velocities of the electrons emitted from a pulsed, activated oxide-coated cathode is much greater than could be expected if the distribution

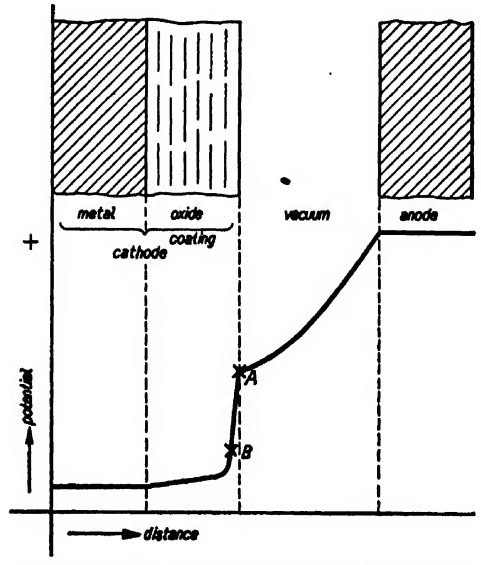


FIG. 1. Potential distribution in an oxide-coated cathode.

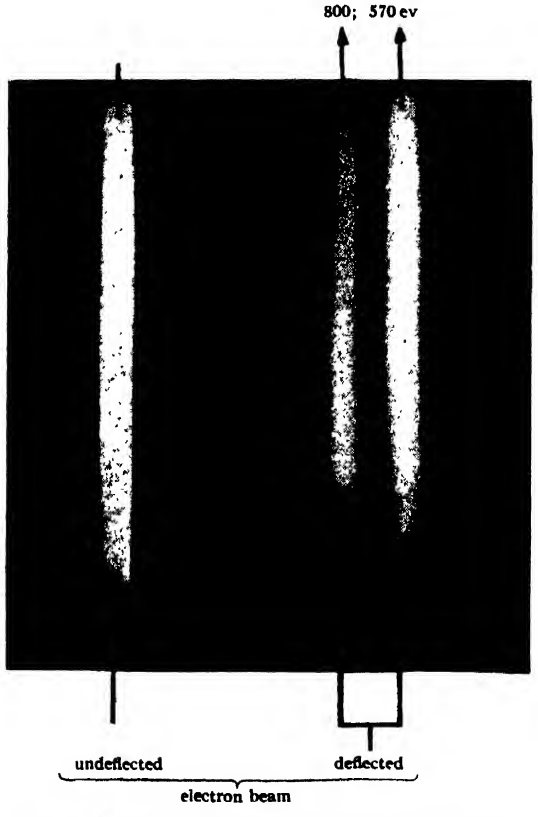


FIG. 2. Electron velocity spectrum of a pulsed oxide cathode. $V_a = 860$ v; $I_a = 14$ amp./cm²; $T = 942$°C.

were Maxwellian. Thus our working hypothesis was confirmed. For example, using square microsecond pulses of 10 amp./cm[2] at 850°C and an oxide coating of 60 micron thickness the energy difference between the fastest and the slowest electrons detectable was 140 ev, the fastest ones still having an energy 60 ev smaller than that which could be computed from the total cathode-anode potential difference.

From all our results it follows that whereas in our cathodes no interface resistance is detectable, a high resistance layer must exist near the surface of the coating. An explanation for this phenomenon might be (a) the fact that the surface is cooler than the bulk of the coating, (b) the fact that the composition of the surface layer is different from that of the bulk,[4] and/or (c) an exhaustion of the electrons in the surface pores by the emission, since according to our theory of the conduction[5] of the oxide coating these same electrons are mainly responsible for the conduction.

We do not think that our results are in contradiction with those of Eisenstein[6] found with a retarding field method.

Another, but unexpected result of our investigations into the electron velocity distribution was the fact that the distribution spectrum was in most cases found to be more or less concentrated in two and even three "lines" instead of being continuous (see Fig. 2). The reason for the occurrence of these "lines" is not yet clear. It might be that electrons are emitted mostly from the surface of the coating and from the "bottom" of the pores in the surface. In view of the high voltages involved even the possibility of emission of secondary electrons from places on the surface, like A, cannot be excluded.

At low temperatures, e.g. 700°C, the spread in the velocities disappears. This can be explained by the fact that with decreasing temperature the thermionic emission falls off more rapidly than the electronic conduction.[5]

* Composition: 52.5 percent by weight BaO, 47.5 percent by weight SrO.

[1] R. Loosjes and H. J. Vink, Philips Res. Rep. **2**, 190–204 (1947).

[2] R. Loosjes and H. J. Vink, to be published in Le Vide.

[3] C. G. J. Jansen and R. Loosjes, unpublished.

[4] A. Eisenstein, J. App. Phys. **17**, 654–663 (1946). H. Huber and S. Wagener, Zeits. f. tech. Physik **23**, 1 (1942).

[5] R. Loosjes and H. J. Vink, Philips Res. Rep. **4**, 449-475 (1949); J. App. Phys. **20**, 884 (1949).

[6] A. S. Eisenstein, J. App. Phys. **20**, 776–790 (1949).

A Sampling of 1949 Books

THE output of new scientific books during 1949 is indeed impressive both in number of books published and in their quality. It seems as though the exposition pent up during the war years is now gradually escaping, and as a result we have an up-to-the-minute survey or treatise in almost every branch of physics. The sampling presented here is a selection from the new books that have come across the Editor's desk during late 1948 and 1949, and is not in any sense a complete list. In each case a few comments are given which may assist the reader in deciding whether or not the book has further interest for him. For easy reference the books are classified as to subject matter.

MATHEMATICS AND MATHEMATICAL PHYSICS

Principles of Mathematical Physics

By WILLIAM V. HOUSTON (new second edition). Pp. 363+xii. McGraw-Hill Book Company, Inc., New York, 1948. Price $5.00.

The purpose of this text is to introduce the student to the more powerful methods of handling physical problems by giving him an opportunity to apply them to simple cases. Designed for intermediate courses, the book is intended to fill the gap between elementary courses in physics and more advanced courses on special subjects. The abundance of problem material forming an integral part of the text is again emphasized, in this second edition, as a principal teaching aid. The student is encouraged to develop the theory for himself by working out these problems, and is urged to depend as far as possible upon his own resourcefulness, insight, and ingenuity.

The chapters on electricity and magnetism have been expanded, and the m. k. s. system of units has been used throughout.

The Mathematical Analysis of Logic

By GEORGE BOOLE (reprinted). Pp. 82. Philosophical Library, New York, 1949. Price $3.75.

This is the first reprinting of a monograph published a century ago and now extremely rare. This short book is believed to have been the first to set on a scientific basis the mathematization of logic and thus prepare the way for the formal study of the foundations of pure mathematics itself.

Elementary Mathematics from an Advanced Standpoint, Vol. II—Geometry

By FELIX KLEIN. Translated from the German by E. R. Hedrick and C. A. Noble. Pp. 214+ix, Figs. 141. Dover Publications, New York, 1949. Price $2.95.

Unlike an ordinary textbook on geometry, this book is intended to give the instructor a comprehensive view of the whole field of geometry, help stimulate his own creative thinking on the subject, and provide him with a broader understanding which he will be able to apply profitably in the classroom, laboratory, or in general reading. The book is divided into three sections: the Simplest Geometric Manifolds, Geometric Transformations, and Systematic Discussion of Geometry and Its Foundations.

Proceedings of Symposia in Applied Mathematics

Pp. 219+vii. American Mathematical Society, New York, 1949. Price $5.25.

This book consists of a series of papers on non-linear problems in the mechanics of continua, including hydrodynamics, elasticity, and plasticity.

Problem Book in the Theory of Functions, Vol. I

By KONRAD KNOPP. Pp. 126. Dover Publications, New York, 1949. Price $1.85.

Used in conjunction with Professor Knopp's previously published *Theory of Functions*, this new English translation of the Problem Book provides the teacher and student with excellent aid in any first course of study covering the theory of functions of a complex variable. This book may also be used to supplement any other textbook in the field.

Introduction to the Theory of Fourier's Series and Integrals

By H. S. CARSLAW (third revised edition). Pp. 368+xiii, Figs. 39. Dover Publications, New York, 1949. Price $3.95.

In preparing this new printing of the third edition, the author has contributed a number of corrections noted since the last printing. This volume contains new tests for uniform convergence of series, a more thorough treatment of term by term integration and the Second Mean Value Theorem, enlarged sets of examples on Infinite Series and Integrals, a new section dealing with the Riemann-Lebesque theorem and its consequences, and a special treatment of the Lebesque definite integral which compares it with the classical Riemann integral.

Modern Operational Calculus

By N. W. MCLACHLAN. Pp. 218+xiv, Figs. 29. The Macmillan Company, New York, 1949. Price $5.00.

Integraltafel. Erster Teil: Unbestimmte Integrale

By WOLFGANG GRÖBNER AND NIKOLAUS HOFREITER. Pp. 166. Springer-Verlag, Vienna, 1949.

Introduction to Applied Mathematics

By FRANCIS D. MURNAGHAN. Pp. 389. John Wiley and Sons, Inc., New York, 1948. Price $5.00.

This is a detailed and self-contained study of the mathematics used in modern physics and engineering. The author gives a relatively full account of vector and matrix calculus, harmonic analysis, spherical harmonics, and Bessel functions. There is a thorough discussion of boundary value problems and integral equations, culminating in Rayleigh's principle; a treatment of mechanical problems by means of the calculation of variations; and a detailed explanation of operational calculus. The book is designed for graduate students and scientific workers with diversified interests.

Numerical Calculus

By WILLIAM EDMUND MILNE. Pp. 393+x. Princeton University Press, Princeton, 1949. Price $3.75.

This book is especially designed for the use of actuarial students, statisticians, applied mathematicians, and others who are forced to seek numerical solutions. It presupposes only a knowledge of the

essentials of algebra, analytical geometry, trigonometry, and elementary calculus. The object is practical, for while numerical calculus is based on the concepts of pure mathematics, it is recognized that the worker must produce a numerical result. This book would be a useful text for undergraduate courses in finite differences and methods of approximation.

Tables of Bessel Functions of Fractional Order, Vols. I and II

By Computation Laboratory of the National Bureau of Standards. Pp. 413 and 365. Columbia University Press, New York, 1948 and 1949. Price $7.50 and $10.00.

These volumes are a continuation of a series published by the Computation Laboratory of the National Bureau of Standards.

Vorträge über Determinanten und Matrizen mit Anwendungen in Physik und Technik

By Werner Schmeidler. Pp. 155. Akademie-Verlag, Berlin, 1949.

Tables of Inverse Hyperbolic Functions

By the Staff of the Computation Laboratory, Harvard University. Pp. 290. Harvard University Press, Cambridge, 1949. Price $10.00.

This is believed to be the first adequate presentation of tables of the inverse hyperbolic functions between the covers of a single volume. The inverse hyperbolic sine, cosine, and tangent are explicitly tabulated to nine decimal places. Differences are available to provide second-order interpolation to this accuracy, and the other inverse hyperbolic functions are readily obtainable. These tables will meet the growing demand in the fields of physics, statistics, engineering, and applied mathematics generally. The book contains the tables, an introduction to the method of calculation, and what is believed to be a complete bibliography of all existing tables of these functions, with critical notes on their range and reliability.

Tables of Generalized Sine- and Cosine-Integral Functions, Parts I and II

By the Staff of the Computation Laboratory, Harvard University. Pp. 462 and 560. Harvard University Press, Cambridge, 1949. Price $10 and $10.

The tables in these two volumes were computed by the automatic sequence controlled calculator, which was presented to Harvard University in 1944 by the International Business Machines Corporation. These tables, which are carried to six decimal places, are invaluable time-savers in the investigation of such questions as self- and mutual impedances, radiation resistance, and distribution of current in antennas and antenna arrays of various types.

Tables of the Bessel Functions of the First Kind of Orders 52 through 63

By the Staff of the Computation Laboratory, Harvard University. Pp. 544. Harvard University Press, Cambridge, 1949. Price $8.00.

Tables of Sines and Cosines to Fifteen Decimal Places at Hundredths of a Degree

By the Staff of the National Bureau of Standards. Pp. 95. Superintendent of Documents, Washington, 1949. Price $0.40.

Tables of Scattering Functions for Spherical Particles

By the Staff of the National Bureau of Standards. Pp. 119. Superintendent of Documents, Washington, 1949. Price $0.45.

The Slide Rule

By Lee H. Johnson. Pp. 242+x, Figs. 138. D. Van Nostrand Company, Inc., New York, 1949. Price $3.00.

This book presents a technique of operation for the Duplex slide rule which makes full use of its versatility and efficiency. A chapter on the Mannheim type is also included.

A Concise History of Mathematics, Vols. I and II

By Dirk J. Struik. Pp. 299+xviii (total), Figs. 49 (total). Dover Publications, New York, 1948. Price $1.50 per volume, $3.00 set.

The author has compressed the history of mathematics into two small volumes. He stresses the continuity and affinity of the Oriental civilizations rather than the mechanical divisions of Egyptian, Babylonian, Chinese, Indian, and Arabian. He is particularly careful to point out the distinctions between established facts, hypotheses, and tradition, especially in Greek mathematics.

Mathematics Dictionary

Glenn James and Robert C. James, editors (revised and enlarged edition). Pp. 432+v. D. Van Nostrand Company, Inc., New York, 1949. Price $7.50.

This dictionary has been designed as a reference work for students and for practicing physicists, chemists, engineers, and statisticians. To the coverage of terms included in the earlier version of the book there have now been added the basic terms in: metric differential geometry, theory of functions of real and complex variables; advanced calculus, differential equations, theory of groups, theory of matrices, theory of summability, point-set topology, analytic mechanics, theory of potential, and statistics. The appendix contains the tables most frequently referred to by mathematicians, physicists, chemists, engineers, and statisticians.

Regular Polytopes

By H. S. M. Coxeter. Pp. 321+xx. Pitman Publishing Corporation, New York, 1949. Price $10.00.

This book attempts to answer questions about the five regular solids. A knowledge of high school mathematics is all that is required for understanding it. There are excursions into such varied subjects as algebra, vector analysis, plane and spherical trigonometry, analytical geometry, integral calculus, crystallography, and topology.

Giant Brains, or Machines That Think

By Edmund C. Berkeley. Pp. 270+xvi. John Wiley and Sons, Inc., New York, 1949. Price $4.00.

Machines that calculate, make decisions, choose between different courses of action, determine most of their own instructions, machines that learn, remember, reason, machines, in other words, that seem almost to think—these machines are the subject of this book. The book describes several existing large-scale mechanical computers, but it is more than a purely descriptive work. Here computing machines are set in perspective; the reader can evaluate their importance to himself and to society. The author compares the machines with human brains and with various types of languages; he indicates their possibilities for the future and the problems of control to direct them toward the greatest good. More than 250 references to the literature are given.

Description of a Relay Calculator

By the Staff of the Computation Laboratory, Harvard University. Pp. 366+xvi. Harvard University Press, Cambridge, 1949. Price $8.00.

This is a description of the relay calculator formerly known as Mark II, that was built by the Computation Laboratory for the

Bureau of Ordnance of the United States Navy. The calculator is now in operation at the Naval Proving Ground, Dahlgren, Virginia. Tape-controlled and completely automatic, the calculator can handle numerical quantities consisting of ten significant digits within the range 10^{-16} to 10^{+16}. It is operated by timed electric impulses provided by a cam unit. All components are duplicated so that the machine may be operated as two independent calculators, solving two problems simultaneously, or as one calculator working on a single problem. The volume includes a complete description of the design and construction of the calculator, a manual of operation, and a discussion of problem preparation and solutions of typical examples. There are 36 plates and many figures illustrating details of construction and wiring.

MECHANICS

Principles of Mechanics

By John L. Synge and B. A. Griffith (second edition). Pp. 530+xvi, Figs. 164. McGraw-Hill Book Company, Inc., New York, 1949. Price $5.00.

This is the second edition of a well-known text in intermediate mechanics. Most readers will be delighted with the philosophical ideas expressed from time to time throughout the book. It concerns the usual chapters in statics, kinematics, dynamics, first considered in a plane and then in a space. Vector methods are used throughout. Three welcome chapters near the end of the book concern Lagrange's equation, special theory of relativity, and theory of dimensions.

Dynamic Principles of Mechanics

By David Rittenhouse Inglis. Pp. 174+x, Figs. 64. The Blakiston Company, Philadelphia, 1949. Price $4.00.

This text is intended for juniors and seniors. Many of the chapters are selected because they illustrate principles fundamental to the understanding of certain important parts of atomic and molecular physics and chemistry. Vector methods are used throughout eight chapters entitled Kinematics of a Particle, Dynamics of a Particle, Coupled Oscillators and Normal Coordinates, Inverse-Square Force and Planetary Motion, A Dynamical System of Several Particles, Statics, Rotation of a Rigid Body, and Impact.

Vectorial Mechanics

By E. A. Milne. Pp. 382+xiii, Figs. 112. Interscience Publishers, Inc., New York, 1948. Price $7.50.

Professor Milne's book is one for advanced students. In his words, he is expounding the view that vectorial mechanics is suitable not only for elegant proofs of general theorems but as a powerful weapon of workaday mathematical investigation, both in research and in solving problems of the type set in English examinations. There are chapters not only on vector analysis but also on tensors. The three main sections of the book deal with vector algebra, systems of line vectors, and dynamics.

Mechanics

By John W. Breneman (new second edition). Pp. 144+xi, Figs. 214. McGraw-Hill Book Company, Inc., New York, 1948. Price $2.50.

The subject covered in this book is usually referred to as engineering mechanics. The text is on a rather elementary level and considers fundamental principles, equilibrium, with examples in bridges and trusses, screws and threads, friction, centroids, curvilinear motion, with a final chapter on work, energy, and power. Helpful tables are provided in appendices concerning trigonometric functions, elements of sections, and so forth.

The Measurement of Stress and Strain in Solids

By the Institute of Physics (London). Pp. 114+x, Figs. 54. Institute of Physics, London, 1948.

This little volume was based on the proceedings of a conference arranged by the Manchester and District Branch of the Institute of Physics in July, 1946. Some of the papers read to the conference were intended to be survey papers, while others were more concerned with the details of the application of the various techniques to specific problems. Twelve papers are included in this book, including three survey papers on electrical resistance strain gauges, recent developments in photoelasticity, and other methods for measurement of strain, such as extensometers and the use of brittle lacquer coatings.

Reissner Anniversary Volume

Polytechnic Institute of Brooklyn, Department of Aeronautical Engineering and Applied Mechanics, editor. Pp. 493+viii. J. W. Edwards, Ann Arbor, Michigan, 1949. Price $6.50.

This volume consists of a collection of contributions to applied mechanics by the friends and former students of Hans J. Reissner on the occasion of his 75th birthday. In addition to the biographical sketch of Dr. Reissner, there are collections of papers in the fields of aerodynamics, dynamics, electricity, elasticity and structures, mathematical methods, plasticity, and propulsion. The reader will find represented in this list many of the great names in mathematical physics.

Elements of Strength of Materials

By S. Timoshenko and Gleason H. MacCullough (third edition). Pp. 426+ix, Figs. 386. D. Van Nostrand Company, Inc., New York, 1949. Price $5.00.

Whereas the original edition of this work presents both the elementary and advanced aspects of the subject, the present work is considerably abridged and is designed primarily for undergraduate courses in elementary strength of materials in engineering colleges.

Applied Mechanics and Statics

By George W. Housner and Donald E. Hudson. Pp. 220+ix. D. Van Nostrand Company, Inc., New York, 1949. Price $3.00.

This book has been prepared as a textbook for the use of sophomore engineering students. The emphasis is on fundamental principles, but the book is not to be considered as a treatise or reference book. The general topics treated are the composition and resolution of force systems, equilibrium, principles of virtual displacement, center of gravity, friction, statically determinant structures, and statically undeterminant structures, with an appendix containing a short bibliography and tables of centroids, moments, and products of inertia.

Mechanics, Statics, and Dynamics

By Merit Scott. Pp. 394+xi. McGraw-Hill Book Company, Inc., New York, 1949. Price $4.50.

The purpose of this text is to serve as a bridge between the first course in college physics and later studies making use of higher mathematics. It is intended also to serve as a reference book, and the initial chapters deal with vector notation and vector analysis. The book is divided into two sections, the first concerning statics and the second, dynamics. Motion is considered, first in one, then in two, and finally in three dimensions.

Elements of Mechanical Vibration

By C. R. Freberg and Emory N. Kemler (second edition). Pp. 227+xiii. John Wiley and Sons, Inc., New York, 1949. Price $3.75.

This is the second edition of the authors' text on mechanical vibration, which concerns primarily the more elementary phases of vibrations and attempts to reduce them to a form in which they can be applied to practical problems. The book is concerned particularly with those problems which can be solved by the use of the simpler forms of differential equations, approximation methods, or the mobility method. Some of the chapter headings are: Vibrations without Damping, Damped Vibrations, Vibration Isolation and Absorption, Equivalent Systems, Beams, Sound, The Mobility Method, and Mechanical and Electrical Models of Vibration Systems.

Mechanical Vibrations

By William Tyrrell Thomson. Pp. 222+viii, Figs. 212. Prentice-Hall, Inc., New York, 1948. Price $5.00.

This text is written primarily for class instruction and follows closely material presented in the course in mechanical vibrations offered to senior and graduate students at the University of Wisconsin. An elementary knowledge of calculus, dynamics, and strength of materials is presupposed, but not familiarity with differential equations. The chapters include: Undamped and Damped Free Vibrations, Forced Vibrations with One, Two, and Many Degrees of Freedom, Vibration of Elastic Bodies, and a chapter on Electromechanical Systems and Analogues.

Design of Machine Elements

By M. F. Spotts. Pp. 402. Prentice-Hall, Inc., New York, 1948. Price $6.65.

This book is intended as a two-semester text for third- and fourth-year engineering students. A brief review of strength of materials is given, in order to emphasize to the student the importance of certain fundamental rules. The chapters include stress concentration and repeated loading, theory of torsion, springs, screws, belts, brakes, clutches, welded and riveted connections, lubrication, ball and roller bearings, spur gears, helical, beveled, and worm gears, and a final very useful chapter on engineering materials.

Analysis and Lubrication of Bearings

By Milton C. Shaw and E. Fred Macks. Pp. 618+xvi. McGraw-Hill Book Company, Inc., New York, 1949. Price $10.00.

The purpose of this book is aptly indicated by a quotation in the preface by A. G. M. Mitchell: "To the machine designer, all bearings are, of course, only necessary evils. . . . Their merit consists in absorbing as little power as possible, wearing out as slowly as possible, occupying as little space as possible, and costing as little as possible." With this philosophy in mind, the general principles of bearing load are developed, and a number of representative examples are discussed. Dimensional analysis is introduced, and there are chapters devoted to hydrodynamic lubrication, rolling contact bearings, boundary lubrication, and a final chapter devoted to a discussion of a representative group of bearing test machines.

Computation Curves for Compressible Fluid Problems

By C. L. Dailey and F. C. Wood. Pp. 33+x, Figs. 33. John Wiley and Sons, Inc., New York, 1949. Price $2.00.

An Introduction to the Mechanics of Viscous Flow

By H. F. P. Purday. Pp. 185+vi, Figs. 79. Dover Publications, Inc., New York, 1949. Price $2.75.

This book is intended as a text to present for the student the application of modern fluid dynamics to problems of lubrication and heat transfer. The author's main object is to show by the simplest means how to calculate velocities, pressures, and boundary resistances arising in laminar fluid motion, and how to adapt the results to serve the theories of bearing lubrication and heat transfer. The book was published in England under the title, *Streamline Flow*.

Deformation and Flow - An Elementary Introduction to Theoretical Rheology

By Markus Reiner. Pp. 346+xix. Interscience Publishers, Inc., New York (U.S. agents), 1949. Price $6.50.

The author explains that this book does not attempt to cover the whole field of rheology. To material mainly of interest to the chemist there has been added other material which should interest the structural or mechanical engineer. The book emphasizes theory, but includes three numerical examples of the application of theory upon one problem each of viscosity, plasticity including strain-hardening, and creep. These were taken from experimental investigations.

The Physics of Rubber Elasticity

By L. R. G. Treloar. Pp. 254. Oxford University Press, New York, 1949. Price $6.00.

This is the fourth volume in a new series entitled "Monographs on the Physics and Chemistry of Materials." The series is intended to summarize the recent results of research in materials and allied subjects, in a form useful to physicists in university and industrial laboratories. The present volume attempts to set down as concisely as possible the essential advances in the study of the elasticity of rubber. The statistical or kinetic theory of elasticity is developed, and is shown to give a satisfactory account of the fundamental mechanical and photoelastic properties of rubber. The book also includes a discussion of the phenomena of crystallization, stress relaxation and flow, the dynamic properties of rubber, and the solution of engineering problems involving large elastic deformations.

Plasticity as a Factor in the Design of Dense Bituminous Road Carpets

By L. W. Nijboer. Pp. 184+xv, Figs. 51. Elsevier Publishing Company, Inc., New York, 1948. Price $5.25.

The importance of a thorough knowledge of the mechanical properties of bituminous road materials is pointed out, and a study is made of the plastic properties of this material. A new test method is developed on the principles of soil mechanics.

HEAT AND THE THEORY OF GASES

Introduction to Statistical Mechanics

By Ronald W. Gurney. Pp. 268+vii, Figs. 59. McGraw-Hill Book Company, Inc., New York, 1949. Price $5.00.

This is an advanced text for graduate students of physics and chemistry. Against a background of atomic physics, the book studies the behavior of large groups of particles, showing how the behavior of matter in bulk follows directly from the properties of the individual atoms. The author has devised methods whereby the student is enabled to grasp the basic principles and the nature of partition functions. A feature of the book is the treatment of metallic alloys, especially the solutions in the alpha- and gamma-phases of iron which form the basis of steel metallurgy. Problems appear at the end of each chapter.

Mathematical Foundations of Statistical Mechanics

By A. I. Khinchin. Translated from the Russian by G. Gamow. Pp. 179+viii. Dover Publications, New York, 1949. Price $2.95.

This book is a mathematical treatment of statistical mechanics on the basis of modern concepts of the theory of probability and a

maximum utilization of its analytical apparatus. The book is written for the mathematician, and its purpose is to introduce him to the problems of statistical mechanics in an atmosphere of logical precision. Among the topics treated in a somewhat novel fashion are the ergodic problems, properties of entropy, and intramolecular correlation.

Scientific Foundations of Vacuum Technique

By Saul Dushman. Pp. 882+xi. John Wiley and Sons, Inc., New York, 1949. Price $15.00.

This book is a broad study of all phases of achieving, maintaining, and measuring very low gas pressures. In the author's words, it presents a survey of "fundamental ideas in physics, chemistry, and (to a smaller extent) metallurgy which will be found useful to both scientists and engineers in dealing with problems in the production and measurement of high vacua."

Vacuum Equipment and Techniques

A. Guthrie and R. K. Wakerling, editors. Pp. 264+xvii. McGraw-Hill Book Company, Inc., New York, 1949. Price $2.50.

This book is concerned with the development and study of high vacuum equipment by the personnel of the University of California Radiation Laboratory for use in electromagnetic separation plants. The problems involved were of a magnitude which made it necessary to do a considerable amount of pioneering work on both equipment and testing. One topic of particular interest is the use of the vacuum analyzer and helium leak detector which were developed for use in the project and proved to be very sensitive instruments.

Glass-to-Metal Seals

By J. H. Partridge. Pp. 238+xii, Figs. 167. The Society of Glass Technology, Sheffield, England, 1949. Price £1 15s. 0d, plus 6d. postage.

This book brings together scattered data about glasses, metals, metal oxides, and the techniques employed in making glass-to-metal seals. It also includes a chapter on the sealing of ceramic materials to metals. The sealing of different glasses to one another and of glasses to ceramic materials are described. The problems of the stresses in the seals and the variation of those stresses when the devices embodying them are in service are also thoroughly examined.

Vacuum Manipulation of Volatile Compounds

By R. T. Sanderson. Pp. 162+viii, Figs. 40. John Wiley and Sons, Inc., New York, 1948. Price $3.00.

This book brings together, in a form designed for easy reference, all the varied essentials which make up the chemical vacuum technique. Through the book this technique becomes readily accessible to the investigator who has no more than the ordinary chemical background. The reader is given a firm foundation for working with high vacuum apparatus and is supplied with all the data necessary for the construction and operation of such apparatus. The book includes vapor pressures, at the same easily producible low temperatures, of almost 400 of the many pure compounds which can conveniently be studied by use of the described manipulative techniques.

A Textbook of Heat

By L. D. Weld. Pp. 436. The Macmillan Company, New York, 1948. Price $5.00.

This is an up-to-date textbook of heat and elementary thermodynamics for upperclassmen. The aim has been to keep the treatment well within the comprehension of students familiar with differential and integral calculus. A carefully selected list of references for supplementary reading is given at the end of each chapter.

Engineering Developments in the Gaseous Diffusion Process

By Manson Benedict and Clarke Williams. Pp. 127+xx. McGraw-Hill Book Company, Inc., New York, 1949. Price $1.25.

This volume describes some of the new engineering procedures and devices developed in connection with the design and construction of gas diffusion plants for separation of uranium isotopes. Topics treated include: special plant instruments and devices including the mass spectrometer for continuous analysis of gas mixtures flowing through a plant, and a recording ionization chamber for detecting traces of radioactive components in a gas mixture; mass spectrometer and engineering methods for testing vacuum-tight systems; engineering processes for handling fluorine and heat transfer ideas applicable to the freezing of liquids.

Heat Transfer

By Max Jakob. Pp. 758+xxix. John Wiley and Sons, Inc. New York, 1949. Price $12.00.

This is a comprehensive work suitable both as a reference work and as a textbook. It shows the development of the ideas which have led to the present knowledge of the subject. The book is international in scope, with particularly thorough coverage of German literature, and should serve the needs of mechanical, chemical, electrical, and civil engineers, and physicists alike.

Third Symposium on Combustion and Flame and Explosion Phenomena

By the Standing Committee on Combustion Symposia. Pp. 748+xiii. The Williams and Wilkins Company, Baltimore, Maryland, 1949. Price $13.50.

This volume contains 100 papers, arranged under eleven categories, which were presented at the Symposium held at the University of Wisconsin in September, 1948. Thirty-eight of the papers were received from European countries.

Steam Power Plants

By Philip J. Potter. Pp. 503+xii. The Ronald Press Company, New York, 1949. Price $6.50.

The aim of this textbook is to present the fundamentals of engineering as applied to the design of steam power plants. No attempt has been made to illustrate all of the various designs of equipment that are available on the market nor to compare similar designs of competitor manufacturers. Descriptions of equipment have been kept to a minimum, and discussions of empirically designed apparatus—such as coal and ash handling systems—have been omitted. Problems and examples have been taken from practical situations to illustrate the principles developed throughout the text.

Thermodynamic Charts for Combustion Processes. Part I Text; Part II Charts

By H. C. Hottel, G. C. Williams, and C. N. Satterfield. Pp. 75, Charts 23. John Wiley and Sons, Inc., New York, 1949. Price $2.60 and $2.40.

ACOUSTICS

Acoustic Measurements

By Leo L. Beranek, Pp. 914+vii. John Wiley and Sons, Inc., New York, 1949. Price $7.00.

In addition to an exposition of acoustic measuring techniques, this book combines data published for the first time with information previously found only in isolated publications. Topics

range from the calibration of microphones and loudspeakers to evaluation of over-all audio systems, and to chapters on the audiometer, speech articulation tests and the sound level meter. The book presents the source of the theory of many electroacoustic phenomena, a description in detail of the basic types of acoustic measuring devices, such as microphones, sources, acoustic impedance bridges and analyzers, and a discussion of the relative advantages and disadvantages of alternate methods for accomplishing electroacoustic measurements.

Sound Absorbing Materials

By C. Zwikker and C. W. Kosten. Pp. 174+ix, Figs. 92. Elsevier Publishing Company, Inc., New York, 1949. Price $3.00.

This book examines the properties that impart sound absorbing qualities to materials or composites, describes their theoretical and experimental study, and shows how these qualities can be conferred upon structures and how they can in many cases be predicted for given surfaces. This latter can be done from mechanical factors such as porosity, compliance, air-resistance, shape, etc. The authors stress the importance of acoustic impedance, which can with advantage be expressed and studied graphically by means of impedance contours. The acoustic behavior of a wide range of materials is discussed and explained.

Theory of Hearing

By Ernest Glen Wever. Pp. 484+xiii, Figs. 136. John Wiley and Sons, Inc., New York, 1949. Price $6.00.

This book is a result of the electrophysiological investigation of the ear's behavior, an investigation begun in 1930 by the author and C. W. Bray and now being carried out actively in many laboratories. The book reviews the development of auditory theory from earliest times and attempts an impartial presentation of all the classical theories of importance. It takes up the principal forms of these theories and brings them face to face with the modern facts about the ear and its actions. It also presents a new point of view in which the two main types of theory—the place and frequency theories—are fused into one and then evaluated with respect to the facts of hearing.

Elements of Sound Recording

By John G. Frayne and Halley Wolfe. Pp. 686+vii, Figs. 480. John Wiley and Sons, Inc., New York, 1949. Price $8.50.

This book covers in detail those subjects which belong peculiarly to the restricted field of sound recording and reproducing and which are not discussed in books devoted to the allied fields of electronics, radio engineering, etc. However, since those interested in studying sound recording have widely different kinds of training and experience, some closely related subjects such as electrochemical analogies, acoustics, vacuum tubes, and audio amplifiers are discussed briefly. For added clarity, numerical examples are used in a number of instances to illustrate the use of design formulas. Complex mathematical analysis is limited to those subjects in which it is desirable for a basic understanding of the point under discussion. The topics covered range from the nature of sound to stereophonic recording.

Installation and Servicing of Low Power Public Address Systems

By John F. Rider. Pp. 208. John F. Rider Publisher, Inc., New York, 1948. Price $1.89.

Inasmuch as the majority of public address installations are those requiring low power equipment—from both an acoustical and economic viewpoint—the apparatus discussed in this book in the main is confined to a rating of 50 watts or lower. Yet a great amount of the subject matter applies equally well to systems of higher power. The book opens with a discussion of the fundamentals of sound and the problems encountered in the public address field; this is followed by an explanation of microphones, phonograph pick-ups, and radio receivers as inputs to an amplifier, then the various types of audio amplifiers, with a chapter devoted to impedance matching.

ELECTROMAGNETIC WAVES AND ELECTRONICS

Principles of Electricity and Electromagnetism

By G. P. Harnwell. Pp. 670+xvi. McGraw-Hill Book Company, Inc., New York, 1949. Price $6.00.

This is a thorough revision of a well-known text for advanced students, and incorporates many important developments in electricity during the past decade. The emphasis is on the phenomena rather than the application of mathematical techniques to particular soluble problems. Increased attention is accorded to the application of the principles of electricity and magnetism to atomic phenomena. Sections have been added on ultra-high frequency techniques, especially those which represent important extensions of the field as covered in the previous edition. No effort has been made to cover the field of radar, but the basic techniques involved are specifically described in this edition. Discussions of wave guides, resonant cavities, and other new topics are included.

Principles of Electricity

By Leigh Page and Norman I. Adams, Jr. (second edition). Pp. 619+xiii, Figs. 323. D. Van Nostrand Company, Inc., New York, 1949. Price $5.50.

This second edition has been brought up to date by the addition of new topics such as the cyclotron, the betatron, and the wave guide, and by the omission of a few topics of diminishing interest. The Gaussian system of units has been employed throughout the book.

Elektrizitätslehre

By R. W. Pohl. Pp. 297. Springer-Verlag, Berlin, 1949.

The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications

By Norbert Wiener. Pp. 163+ix. John Wiley and Sons, Inc., New York, 1949. Price $4.00.

This book represents an attempt to unite the theory and practice of two important fields of work—time series in statistics, and communication engineering. These fields are widely different in vocabulary and in the training of personnel, and they have drawn their inspiration from two entirely distinct traditions. However, it is believed that they have a unification of theory and method. The author points out that the statistician is completely familiar with methods involving probability theory and correlation, but unable to use the engineer's methods; the communication engineer is just as familiar with the use of the complex plane, but unable to use the statistician's techniques. It is the purpose of this book to introduce methods stemming from both of these existing techniques and to fuse them into a common technique which is more effective than either one alone.

Electric and Magnetic Fields

By Stephen S. Attwood (third edition). Pp. 475+xi. John Wiley and Sons, Inc., New York, 1949. Price $5.50.

This book was written to provide a smooth transition from the study of mathematics, mechanics, and physics to advanced, professional-level electrical engineering. The author has divided this new edition into four parts. The electric field is discussed first. Next the magnetic field is treated and the field formulas are devel-

oped under the assumption of constant permeability. In the third part the ferromagnetic field introduces the modification of the basic magnetic relations necessary to account for the use of magnetic material in which permeability cannot be considered constant. The fourth section treats in an elementary manner the interactions between electric and magnetic fields that simultaneously occupy the same space. The formulas in the book have been recast in the rationalized form of the m.k.s. system.

Basic Electrical Engineering

By George F. Corcoran. Pp. 449+vii. John Wiley and Sons, Inc., New York, 1949. Price $4.50.

This is a textbook of electrical engineering which maintains throughout a level suitable for beginners. Each electrical or magnetic concept is exactly defined at the place of its introduction. Arrangement of subject matter is likewise planned to suit the beginning course. First d.c. theory is considered, with d.c. circuit analysis by the mesh and nodal methods. This is followed by electrical field theory, magnetic field theory, inductance, capacitance, and non-linear circuit elements. There is a wide range of problems with each chapter.

Electronics in Engineering

By W. Ryland Hill. Pp. 274+viii. McGraw-Hill Book Company, Inc., New York, 1949. Price $3.50.

This text is designed for senior and graduate students in engineering and allied sciences who find the need for an understanding of electronics and its applications to measurement and control in their respective fields. The treatment is sufficiently analytical to interest such students, and yet contains enough practical applications to illustrate the uses of electronics to the various engineering fields. The book provides material for a one-quarter or one-semester course covering vacuum and gas-filled tubes, rectification, amplification, feedback, the cathode-ray oscilloscope, and simple control and instrumentation techniques.

Electrical Engineers' Handbook; Electric Power. Vol. I

Harold Pender and William A. Del Mar, editors (fourth edition). Pp. 1700+xii. John Wiley and Sons, Inc., New York, 1949. Price $8.50.

The editors point out that the growth of knowledge and the greater degree of specialization in the various phases of electrical engineering have necessitated a considerable enlargement of this well-known handbook. An effort has been made to keep the book compact and readable.

Learning Electricity and Electronics Experimentally

By Leonard R. Crow. Pp. 525+xi. Universal Scientific Company, Vincennes, Indiana, 1949.

The author of this book aims to use the experimental or project approach to the subject of electricity, employing the simplest possible language. Some of the specific points treated in the book are energy and electrical effects, magnetism and electromagnetism, electromagnetic induction, magnetic flux control, rotating magnetic fields, current control by magnetic saturation.

Theory of Oscillations

By A. A. Andronow and C. E. Chaikin. Pp. 358+ix, Figs. 313. Princeton University Press, Princeton, 1949. Price $6.00.

This is a condensed and edited translation of the first comprehensive treatment of non-linear oscillations ever to be published. The original work, by two leading members of the Soviet Institute of Oscillations, was published in Russian in 1937. Since the field of non-linear oscillations is not as well known as it deserves to be, numerous examples are given, and a special index of physical examples is included. Practical applications are emphasized, and the examples given in the original Russian edition are preserved. The book begins with a treatment of linear systems, giving the newcomer to the field familiar ground from which to work, but it rapidly progresses to various types of non-linear systems, using both mechanical and electrical examples. The book will be useful to electrical, aeronautical, and mechanical engineers, to mathematicians and physicists, and to any scientists working with oscillatory phenomena.

Micro-Waves and Wave Guides

By H. M. Barlow. Pp. 122+x, Figs. 70. Dover Publications, New York, 1949. Price $1.95.

The first part of this volume presents a physical picture of wave-guide modes, synthesized from constituent plane waves in association with ordinary transmission line elements. The coaxial cable, representing the common meeting ground of wave-guide modes and the simple transverse electromagnetic wave, is discussed. The final section deals with the measurements and applications of microwaves, giving the reader an appreciation and knowledge of the methods adopted in actual practice.

Radar Systems and Components

By the Technical Staff of Bell Telephone Laboratories. Pp. 1042+xiv. D. Van Nostrand Company, Inc., New York, 1949. Price $7.50.

About one-half of Bell Telephone Laboratories' total research effort during the war was devoted to radar. The Laboratories shared with the Radiation Laboratory at Massachusetts Institute of Technology the responsibility for the bulk of research and development in this field in the United States. In order that the scientific advances achieved might be recorded in detail (so far as security considerations would permit), the papers appearing in this volume were prepared and were originally published in the *Bell System Technical Journal*. While the work described was directed toward radar development as its immediate objective, it has a broad application to the new field of microwave transmission.

Aerials for Metre and Decimetre Wavelengths

By R. A. Smith. Pp. 218+xii, Figs. 125. Cambridge University Press, New York, 1949. Price $3.75.

The author of this book points out that the extension of radio technique to extremely short wave-lengths greatly increased the scope of aerial design. Half-wave and full-wave dipoles have manageable dimensions when the wave-length is less than 12 meters, and arrays for many special purposes may be designed. In this book the upper limit has been set at 12 meters, the longest wavelength used for radar. The lower limit has been fixed at 10 cm. A limited number of applications have been selected to illustrate general principles.

Electrical Transmission of Power and Signals

By Edward W. Kimbark. Pp. 461+xi. John Wiley and Sons, Inc., New York, 1949. Price $6.00.

This is the first textbook to apply transmission-line theory to all three fields of power, telephone, and ultra-high frequency transmission. It contains transmission-line (hyperbolic function) charts, tables and graphs of the characteristics of power-line conductors, power transmission and distribution lines, telephone lines and cables, submarine telegraph cables, and radiofrequency lines. This volume is a useful reference book for professional engineers and will be found of value in electrical transmission courses.

Elements of Electromagnetic Waves

By Lawrence A. Ware. Pp. 203+x. Pitman Publishing Corporation, New York, 1949. Price $3.50.

This text was prepared to meet the need for an elementary introduction to the basic ideas of electromagnetic theory, employ-

ing the rationalized m.k.s. units. It is intended as first-course material for either juniors or seniors, and the only prerequisites are calculus and the fundamentals of alternating-current theory. The text is prepared to serve as an introduction to the more advanced treatments of the subject.

Radio-Frequency Heating Equipment

By L. L. Langton. Pp. 196+ix, Figs. 95. Pitman Publishing Corporation, New York, 1949. Price $3.75.

This book deals with the design of equipment for radiofrequency heating, and also with some applications of the technique. The reasons underlying the advantages gained by using the technique in these applications are fully discussed; this serves as a guide to the advantages to be expected when the technique is applied elsewhere.

Terrestrial Radio Waves; Theory of Propagation

By H. Bremmer. Pp. 343+x, Figs. 91. Elsevier Publishing Company, Inc., New York, 1949. Price $6.75.

This book describes the mathematical methods by which the electromagnetic field of a radio transmitter can be computed, taking into account the curvature of the earth's surface. Special attention is paid to the physical interpretation of the mathematical deductions. The complicated formulas which are basic for numerical computations of the "ground-wave" have been collected in a practical form in a separate chapter.

Earth Conduction Effects in Transmission Systems

By Erling D. Sunde. Pp. 373+xiii. D. Van Nostrand Company, Inc., New York, 1949. Price $6.00.

This presentation is primarily concerned with fundamental methods in the analysis of earth conduction effects and basic principles underlying protective measures against resultant circuit disturbances. For conciseness and better exposition of basic formulas and considerations, much of the mathematical detail has been omitted in most derivations or covered by reference to other literature.

The Mathematics of Circuit Analysis

By Ernest A. Guillemin. Pp. 590+xiv. John Wiley and Sons, Inc., New York, 1949. Price $7.50.

This volume contains a collection of a variety of principles and methods essential to a thorough understanding of electrical network theory. It is designed specifically as a source book to supplement the usual undergraduate engineering mathematics curriculum. It is appropriate also for use as a textbook in a course in higher mathematics for engineers.

Communication Circuits

By Lawrence A. Ware and Henry R. Reed (third edition). Pp. 403+x. John Wiley and Sons, Inc., New York, 1949. Price $5.00.

In this third edition the authors cover the essentials of the transmission and lead logically to such subjects as filters, impedance matching, and wave guides. Features of this edition include the use of rationalized m.k.s. system of units; a radical revision of the chapter on transmission-line parameters; a treatment of attenuation in wave guides based on Poynting's theorem; the addition of many new problems. Although a knowledge of calculus and the elements of a.c. theory has been assumed, the more advanced mathematics that is needed for certain portions has been included in the appendices.

Pulses and Transients in Communication Circuits

By Colin Cherry. Pp. 317+xvi, Figs. 129. Chapman and Hall, Ltd., London, 1949. Price 32s. net.

This book is meant as an introduction to circuit transient analysis for communications engineers, which will bridge the gap between simple conventional alternating current theory and operational methods of analysis. It is intended to give students of steady-state circuit theory the essential groundwork, using, wherever possible, rigorous physical arguments and only elementary mathematics. Mathematical language and notation are avoided as much as possible. Electrical wave forms are dealt with, rather than analytical functions, thus giving the book a geometrical flavor.

Networks, Lines and Fields

By John D. Ryder. Pp. 462+x, Figs. 195. Prentice-Hall, Inc., New York, 1949. Price $5.50.

In this text the author has attempted to expand upon the subjects of networks, resonance, and wave filters in such a way as to develop in the undergraduate student a definite familiarity with "network theory," or the description of network performance by certain measured or defined parameters. The rationalized m.k.s. system of units is used throughout.

Radio Engineering

By E. K. Sandeman. Pp. 775+xxiv. John Wiley and Sons, Inc., New York, 1948. Price $6.50.

Originally planned as instructions for maintenance engineers at B.B.C. transmitting stations, this book is designed to give beginners in radio the working principles they need. Experienced engineers and designers of radio equipment and circuits will find it a useful reference volume. Practical ends are stressed, either in the form of technical facts, designs, or methods or else in the form of essential formulas.

Electric Circuits and Machines

By B. L. Robertson and L. J. Black. Pp. 434+vii. D. Van Nostrand Company, Inc., New York, 1949. Price $5.00.

This is a textbook intended primarily for third-year students of mechanical engineering, or for non-electricals in general, to give them a strong minor in electrical engineering.

Radio at Ultra-High Frequencies, Vol. II

By Alfred N. Goldsmith *et al.*, editors. Pp. 485. RCA Review, RCA Laboratories Division, Princeton, New Jersey, 1948. Price $2.50.

This is the eighth volume in the RCA Technical Book Series and the second on the general subject of radio at the higher frequencies. The present volume covers the period 1940–47. Papers are presented in seven sections: antennas and transmission lines; propagation; reception; radio relays; microwaves; measurements and components; navigational aids. The appendices include a bibliography in the field of ultra-high frequency radio, and summaries of all papers appearing in Vol. I, which is now out of print.

Electrical Measurements in Theory and Application

By Arthur Whitmore Smith. Pp. 371, Figs. 250. McGraw-Hill Book Company, Inc., New York, 1948. Price $4.25.

This is a thorough revision and modernization of a well-known standard text, incorporating many changes to clarify the explanations and present improved methods. The meter-kilogram-second (m.k.s.) system has been used for the definitions of the units of the electrical and magnetic quantities, resulting in great simplification of the approach to these definitions. Several new measurement methods have been added, especially for the measurement of capacitance and inductance.

Advances in Electronics, Vol. I

L. MARTON, editor. Pp. 475+xi. Academic Press, Inc., New York, 1949. Price $9.00.

This volume is the first of a series of yearly publications devoted to articles in the general field of electronics. The aim is to publish critical and integrated reviews on physical and engineering electronics. The editors and publishers of the book were obliged to use a somewhat arbitrary definition of the word "electronics," and thus limit the choice of materials to a reasonably restricted field. In this volume the topics covered are: oxide-coated cathodes, secondary electron emission, television pick-up tubes and the problem of vision, the deflection of beams of charged particles, modern mass spectroscopy, particle accelerators, ionospheric research, cosmic radio noise, propagation in the f-m broadcast band, electronic aids to navigation.

Electronics: Experimental Techniques

BY W. C. ELMORE AND MATTHEW L. SANDS. Pp. 417+xviii. McGraw-Hill Book Company, Inc., New York, 1949. Price $3.75.

This volume describes the most useful and the most successful circuits for making nuclear and other physical measurements. In addition to the description of complete circuits—pulse amplifiers, counting circuits, sweep circuits, timing circuits, testing equipment, etc.—the book contains a discussion of the circuit elements that constitute the basis of the complete circuits.

Theory and Design of Electron Beams

BY J. R. PIERCE. Pp. 197+xiii. D. Van Nostrand Company, Inc., New York, 1949. Price $3.50.

This book represents an effort to collect the minimum amount of theoretical material necessary for a good understanding of electron flow and electron focusing in devices other than electron microscopes and image tubes. Much of the elementary material found in books on electron optics is included. There is, however, no extensive treatment of aberrations. Instead, there is considerable emphasis and material on space charge and thermal velocities. Problems are included, partly so that the book can be used as a text, but more so that "the reader will have to do things for himself in order to understand them."

Fundamentals of Radio-Valve Technique

BY J. DEKETH. Pp. 535+xxii, Figs. 384. N. V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands, 1949. Price $5.00.

In this book the author aims at giving technicians and research workers an account which will be of value to those who make, use, or maintain apparatus employing electronic valves. The mathematical treatment is restricted, and the subjects treated include resonance, short-wave technique, and low frequency modulation.

Microwaves and Radar Electronics

BY ERNEST C. POLLARD AND JULIAN M. STURTEVANT. Pp. 426+vii. John Wiley and Sons, Inc., New York, 1948. Price $5.00.

No extensive background in communications or electronics is required for use of this book. The usual student background in physics, chemistry, and electrical engineering is sufficient. The essential facts about microwave circuits are developed on the fundamental basis of electricity and magnetism, rather than from an extension of the theory of circuits with lumped parameters. Pulse circuits are considered as a unified subject, since the fundamental problems and basic methods of solving such problems accompany all types of pulse circuits. Much attention has been devoted to the applications of microwave and radar electronic techniques which have been considered in various scientific periodicals since the end of the war.

Industrial Electronics Reference Book

BY ELECTRONICS ENGINEERS OF THE WESTINGHOUSE ELECTRIC CORPORATION. Pp. 680. John Wiley and Sons, Inc., New York, 1948. Price $7.50.

This reference book presents the basic theory and the vital application principles of industrial electronics. It is intended to help the reader to appraise accurately the possibilities as well as the limitations of new electronic devices, and also to help technicians to develop new applications and projects which will have a better chance of giving the desired results. The authors deal mostly with basic electronics application information and treat the specific apparatus less fully because of the many changes still taking place. The generous use of illustrations in the volume helps to elucidate every phase of the text material.

The Characteristics of Electrical Discharges in Magnetic Fields

A. GUTHRIE AND R. K. WAKERLING, editors. Pp. 376+xviii. McGraw-Hill Book Company, Inc., New York, 1949. Price $3.50.

This volume was prepared in an effort to collect the results of studies carried out in the California Radiation Laboratory on the characteristics of electrical discharges in magnetic fields. The primary emphasis is placed on the case of electrical discharges in the vapors of uranium compounds. The book does not cover all of the investigations, although the most significant studies are included. More than half of this book is devoted to work carried out by members of a British scientific mission.

Basic Electronics

BY ROYCE G. KLOEFFLER AND MAURICE W. HORRELL. Pp. 435+xiv. John Wiley and Sons, Inc., New York, 1949. Price $5.00.

The authors cover the fundamental theory of electronics in condensed form, assuming the reader has a knowledge of elementary physics, but no previous training in electronics. The authors' approach covers both the power and communication fields. Included among the subjects discussed are the developments of the transistor and mechano-electronic transducer. As a convenience to the reader, the authors have included curves and data of the basic tubes. This means that most of the problems can be solved without referring to any of the tube handbooks.

Photoelectricity and Its Application

BY V. K. ZWORYKIN AND E. G. RAMBERG. Pp. 494+xii. John Wiley and Sons, Inc., New York, 1949. Price $7.50.

This volume replaces the second edition of *Photocells and Their Application* by V. K. Zworykin and E. D. Wilson. It retains the aim of its predecessor—to present the reader with practical reliable data on the properties, preparation, and applications fo photoelectric devices. Throughout, the emphasis is on practical aspects.

Theory of Dielectrics

BY H. FRÖHLICH. Pp. 180+vi, Figs. 47. Oxford University Press, New York, 1949. Price $4.50.

This book, which is mainly written for applied scientists, is intended to give a systematic account of the theory of dielectric properties, and thus to provide a background against which applications can be made. The reader is assumed to have a certain ele-

mentary knowledge of atomic and molecular physics, statistical mechanics, electrostatics, and little mathematics beyond the calculus.

Atmospheric Electricity

By J. Alan Chalmers. Pp. 175, Figs. 36. Oxford University Press, New York, 1949. Price $3.75.

This book is intended to serve both as an introduction to the study of atmospheric electricity and as a short work of reference. It was written primarily for physicists but should prove useful to meteorologists and others interested in the subject. There are chapters on the ions in the atmosphere, the earth's vertical field, the conductivity of the air, the year-around current, point discharge currents, precipitation currents, the transfer of charge, the thunder cloud, the lightning flash, and the separation of charge.

TV Picture Projection and Enlargement

By Allan Lytel. Pp. 179+xi. John F. Rider Publisher, Inc., New York, 1949. Price $3.30.

This book, with its numerous illustrations, will appeal to those interested in television servicing. The author discusses the fundamentals of light, the Schmidt projection system, and refractive projection systems.

Understanding Television

By Orrin E. Dunlap, Jr. Pp. 128. Greenberg, Publisher, New York, 1948. Price $2.50.

This little book discusses the steps that led to television, the mechanisms which match the human eye, the necessity for high antennas and relay stations, the coaxial cable, and the receiver itself. The author describes how a television program is put together, both in the studio and in the field.

OPTICS AND SPECTROSCOPY

Introduction to Theoretical and Experimental Optics

By Joseph Valasek. Pp. 454+x. John Wiley and Sons, Inc., New York, 1949. Price $6.50.

This is a four-part book, covering geometrical optics, physical optics, radiation and spectra, and experiments relating to each of these fields. In line with the author's special interests, the subjects of x-rays, photographic optics, and ophthalmic lenses are treated in greater detail than is customary in books dealing with optics. This book holds useful information for many engineers and physicists working in industry and in research laboratories. It serves as an introduction to optical engineering, a field which is steadily growing more important. For example, students of photography, mechanical engineers who use optical methods of stress analysis, and optical mineralogists all can use the principles of optics contained in this book. Although the author assumes the reader has a knowledge of calculus and general physics, he keeps the mathematical treatment as elementary as possible.

Optik

By R. W. Pohl. Pp. 356, Figs. 565. Springer-Verlag, Berlin, 1948.

Physical Aspects of Color

By P. J. Bouma. Pp. 312, Figs. 113. Elsevier Publishing Company, Inc., New York, 1947. Price $5.50.

The physiologist, the psychologist, the physicist, the illumination engineer, and the chemist each has studied the problem of color perception from his own point of view and expressed the results in his own particular technical language. The author of this book has aimed to cast the subject matter in such a form that it may well enrich the knowledge of many a scientist and at the same time serve as a guide to the beginner—presumed to possess a secondary school knowledge of mathematics and physics—for the study of problems of chromatics. Starting from the established laws of color mixing, the author gives a complete insight into the origin and perception of colors in a logical and systematic manner.

Elementary Photography

By Guilford G. Quarles (new second edition). Pp. 345+xii, Figs. 86. McGraw-Hill Book Company, Inc., New York, 1949. Price $4.50.

As before, this text provides a treatment of photography between very elementary ones and those of an advanced, technical nature. The book assumes no previous knowledge of the subject, and an effort has been exerted to make the new edition even more easily understandable to the student and more teachable from the point of view of the instructor. A major revision of the order of the chapters has been made in order to bring the student as quickly as possible to active photography—i.e., the making of pictures.

Die Optik in der Feinmehtechnik

By Kurt Räntsch. Pp. 317, Figs. 299. Carl Hanser-Verlag, Munich, 1949.

Photographic Optics

By Arthur Cox. Pp. 412+xi. Pitman Publishing Corporation, New York, 1949. Price $5.75.

The subtitle of this book is "The Technique of Definition." The photographer should be able to discriminate between lenses in the same way he does between films and filters. This practical book, written in simple language and containing many diagrams and tables, deals with the importance of lenses in obtaining the best results with the camera, and shows how the proper type of lens must be used in the right kind of camera for each kind of subject. The photographer is shown why the correct focal length, best aperture, and sufficient covering power are considerations that eventually decide the picture.

Spectroscopic Properties of Uranium Compounds

By G. H. Dieke and A. B. F. Duncan. Pp. 290+xviii. McGraw-Hill Book Company, Inc., New York, 1949. Price $2.75.

This volume represents a comprehensive study of the spectroscopic properties. Correlation of absorption measurements with those obtained from fluorescence spectra have seldom been made for such a large amount of data. This work should be an important source of information and data in the development of future theories of the solid state.

Practical Spectroscopy

By George R. Harrison, Richard C. Lord, and John R. Loofbourow. Pp. 605+xiv. Prentice-Hall, Inc., New York, 1948. Price $6.65.

In this book the authors have avoided the historical method of approach and instead have attempted to give a comprehensive view of the status and possibilities of experimental spectroscopy as it exists today. In Chapter I they "view the field as from a great altitude, to enable the reader who is unacquainted with the methods and accomplishments of spectroscopy to judge for himself which parts, if any, may be of importance to him." In the remainder of the book they reconsider the various topics in considerably greater detail.

Practical Spectroscopy

By C. Candler. Pp. 190+viii. The Jarrell Ash Company, Boston (U.S. agents), 1949. Price $6.10.

This book is intended as a course of instruction in all branches branches of practical spectroscopy. The subject matter has been written primarily around the Hilger Barfit wave-length spectrometer. However, it is believed that this should not detract from its interest to those equipped with other apparatus. Experiments are given at the close of each chapter.

Absorption Spectrophotometry

By G. F. Lothian (third edition). Pp. 196, Figs. 71. Hilger and Watts, Ltd., London, 1949. Price $7.60.

Although this work forms the third edition of an already well-known publication, *The Practice of Absorption Spectrometry*, it has been so completely revised and largely rewritten that it is virtually a new work. It is a valuable handbook for the investigator who is employing spectrophotometry and contains much that will be of interest to those who are contemplating the use of that method. It describes the apparatus to be used and also shows the accuracy to be anticipated and the essential conditions for realizing that accuracy. Examples of work done in selected fields are included.

Infrared Determination of Organic Structures

By H. M. Randall, Nelson Fuson, R. G. Fowler, and J. R. Dangl. Pp. 239+v, Plates 354. D. Van Nostrand Company, Inc., New York, 1949. Price $10.00.

This volume is intended to supply some of the background which is required by investigators, whether they are chemists, biochemists, or physicists, using infra-red spectra to determine molecular structure. Methods of approach have been outlined; the possibilities inherent in the method have been explored; currently established correlations between structure and the absorption spectrum have been itemized; the instruments and techniques have been described.

An Introduction to Molecular Spectra

By R. C. Johnson. Pp. 296+xiii, Figs. 151, Plates 8. Pitman Publishing Corporation, New York, 1949. Price $7.50.

The author has attempted to provide a textbook on molecular spectra for the student who has graduated in physics or chemistry, or who is approaching graduation, and desires an understanding of the subject within the limits of ordinary mathematical equipment at that stage. At one or two points the author has introduced the results of wave mechanics, but has not assumed familiarity in the use of this method.

Contribution to the Study of Molecular Structure

In both French and English. Several contributors, in commemoration of Victor Henri. Pp. 314. Maison Desoer Editions, Liege, Belgium, 1948.

Luminescent Materials

By G. F. J. Garlick. Pp. 254+viii, Figs. 127. Oxford University Press, New York, 1949. Price $5.50.

This book describes recent research on the physical properties of luminescent materials. It is concerned with experimental and theoretical studies of such phenomena as fluorescence, phosphorescence, thermoluminescence, and the effects of infra-red radiation on phosphors. Associated phenomena, such as photoconductivity, dielectric changes, and the action of strong electric fields are also discussed. The preparation, constitution, and classification of materials according to their luminescence characteristics are given an important place in the text.

ATOMIC AND NUCLEAR PHYSICS

Foundations of Nuclear Physics

With bibliography compiled by Robert T. Beyer. Pp. 272. Dover Publications, New York, 1949. Price $2.95.

Here are thirteen important papers on nuclear physics, presented unabridged and in the original language of their authors. Selected from scientific journals published both in this country and throughout the world, these are the essential primary source materials in this field. This volume also contains a comprehensive guide (120 pages in length) to published material on nuclear physics in which entries are classified according to subject matter.

Wave Mechanics and Its Applications

By N. F. Mott and I. N. Sneddon. Pp. 393+xii, Figs. 68. Oxford University Press, New York, 1948. Price $10.00.

The aim of this book is to give a straightforward account of wave mechanics designed for those who want to use it. There are chapters dealing with applications to atomic and molecular structure, radiation, and collision problems, and the properties of solids. Modern relativistic quantum mechanics is also discussed. The book should be useful as a textbook to the theoretical physicist beginning research in the subject, and to his experimental colleagues as a reference book, while several chapters could usefully be studied by an advanced student of physics in his final year.

Theory of Atomic Nucleus and Nuclear Energy-Sources

By G. Gamow and C. L. Critchfield. Pp. 344+vi, Figs. 62. Oxford University Press, New York, 1949. Price $9.00.

This is the third edition of Gamow's *Structure of Atomic Nucleus and Nuclear Transformations*, of which the second edition has been out of print since 1944. The book has been practically rewritten. The authors have attempted to write it in an inductive way, starting from the properties of elementary particles from which the nuclei are made and working up to more and more complex nuclear properties. Considerable space has been devoted to thermonuclear reactions and their application to cosmological problems.

Two Lectures: I. The Present Situation in the Theory of Elementary Particles. II. Electron Theory of Superconductivity

By W. Heisenberg. Pp. 52. Cambridge University Press, New York, 1949. Price $1.00.

The two lectures here printed were delivered at the Cavendish Laboratory as an introduction to discussions on two different topics in atomic theory. The first lecture tries to explain why the well-known divergencies in meson theory and nuclear physics may be considered, not as a difficulty, but rather as a natural and therefore satisfactory feature of the present "correspondence" theory. The second lecture deals with an application of quantum mechanics, the theory of superconductivity. The views presented in it may be considered as an attempt to combine different features of the previous rather discordant theories into a consistent picture of the phenomenon.

The Physical Principles of the Quantum Theory

By Werner Heisenberg. English translation by Carl Eckart and Frank C. Hoyt (reprint of 1930 edition). Pp. 184+viii, Figs. 19. Dover Publications, New York, 1949. Price $2.50.

In this book the author, a Nobel prize winner, gives a physical picture of quantum theory, covering not only his own contribu-

tions but also those of Compton, Simon, Schrödinger, Einstein, Wilson, Rupp, Davisson, Germer, Thomson, Bohr, and others. Although there is a mathematical appendix, only elementary formulas and calculations have been used in the text itself, so far as practical.

Nuclear Radiation Physics

By R. E. Lapp and H. L. Andrews. Pp. 487+xiv. Prentice-Hall, Inc., New York, 1948. Price $6.00.

This volume is the outgrowth of an elementary manual originally written by the present authors and published within the Military Establishment as "Radiological Safety." Basically, the authors have tried to present a logical and simple interpretation of the phenomena of nuclear physics. In each chapter illustrative examples are given in order to serve as an introduction to the calculations of radioactivity and to guide the student in the technique of solving the problems that appear throughout the text. An essentially non-mathematical approach has been employed, in the hope that the subject matter would not be too involved for non-physicists.

The Theory of Groups and Quantum Mechanics

By Hermann Weyl. Pp. 422+xvii. Dover Publications, New York, 1949. Price $4.50.

The author of this book postulates the importance of the theory of groups in discovering the general laws of quantum theory. Showing how concepts arising in the theory of groups find their application in physics, he illustrates his argument with many examples. Included are discussions of Schrödinger's wave equation, de Broglie's waves of a particle, the Jordan-Holder theorum and its analogs, Burnside's theorem, Lie's theory of continuous groups of transformations, the Pauli exclusion principle and the structure of the periodic table, quantization of the Maxwell-Dirac field equations, and Young's symmetry opreators.

The Theory of Atomic Collisions

By N. F. Mott and H. S. W. Massey (second edition). Pp. 388+xv, Figs. 69. Oxford University Press, New York, 1949. Price $8.75.

This new edition contains a systematic treatment of the new "dispersion" method developed for nuclear collisions and of its application to a wide variety of nuclear phenomena—the scattering and radiation capture of neutrons and protons by various nuclei as well as the scattering of alpha-particles. Nuclear fission is also discussed.

Atoms in Action

By George Russell Harrison (third edition). Pp. 406+x. William Morrow and Company, New York, 1949. Price $5.00.

This "layman's guide to modern science," an account of our most important physical discoveries and their contributions to everyday life, has been completely revised and expanded. It includes new material on atomic energy, radar, and electronics. As in previous editions, the author explains difficult technical matters in simple, graphic language.

Elementary Modern Physics

By Gordon Ferrie Hull. Pp. 503+xvi. The Macmillan Company, New York, 1949. Price $5.25.

This book is intended to gather together, for the use of students, the main developments in physics in recent years—to present the experimental facts and the more elementary theories.

Artificial Radioactivity

By P. B. Moon. Pp. 102+x, Figs. 28. Cambridge University Press, New York, 1949. Price $2.50.

This monograph gives an outline of the main phenomena and techniques of radioactivity, as met with in the study of the light-weight and medium-weight radioactive nuclei that can now be made in great variety. It is chiefly intended for readers who wish to be put into touch with the recent literature on the subject. The radioactive properties of a good many individual nuclei are discussed.

The Science and Engineering of Nuclear Power, Vol. II

Clark Goodman, editor. Pp. 317. Addison-Wesley Press, Boston, 1949.

Whereas Vol. I of this work was primarily a general survey of the fundamentals of chain-reacting systems, the present volume considers specific aspects in some detail such as source materials, isotope separation, vacuum techniques, pile kinetics, rockets, health physics, and future developments. The contents should be of value to all scientists and engineers seriously interested in the industrial applications of nuclear energy. The editor states that the contents of these two volumes "alone should be sufficient proof that 'atomic secrets' are largely figments formed in the minds of well-meaning persons who happen to be ignorant of this special field. A workable basis for the free exchange of ideas must be found. Mankind cannot continue to stand in mortal fear of the greatest scientific accomplishment of our time."

Atomic Energy Levels, Vol. I

By Charlotte E. Moore. Pp. 309+xliii. Circular 467 from the National Bureau of Standards, Washington, D. C., 1949. Price $2.75.

This volume is a compilation of all known data on the energy levels of elements of atomic number 1 through 23. It was designed to meet the needs of workers in nuclear and atomic physics, astrophysics, chemistry, and industry. This volume is the first of a series being prepared by the Bureau as part of a general program on atomic energy levels derived from observations of the optical spectra of atoms and ions. Spectra are presented in order of increasing atomic number, and under a given atomic number they are listed in order of increasing stages of ionization. For each spectrum a selected bibliography covering the analysis is given.

Excited States of Nuclei

By S. Devons. Pp. 152+ix, Figs. 27. Cambridge University Press, New York, 1949. Price $2.50.

This monograph was written mainly from the experimental standpoint. It gives a description of experimental methods and examines typical experimental results in detail. An attempt has been made to include examples taken from recent research and particularly to provide a broad picture of the situation which is up to date.

Trilinear Chart of Nuclear Species

By William H. Sullivan. John Wiley and Sons, Inc., New York, 1949. Price $2.50.

This publication was originated in the Chemistry Division of Clinton Laboratories during November 1946 and sponsored initially by it and by the Argonne National Laboratory. Its preparation was also facilitated and supported by the AEC and by the National Research Council.

Ionization Chambers and Counters: Experimental Techniques

By Bruno Rossi and Hans H. Staub. Pp. 243+xviii. McGraw-Hill Book Company, Inc., New York, 1949. Price $2.25.

Part I of this volume discusses the general physical principles underlying the operation of ionization chambers and counters, with special emphasis on subjects that are of importance in modern detection techniques. Original results are included which were obtained at the Los Alamos Laboratory in the course of assembling important nuclear data. Part II presents detailed discussions for a number of detectors developed and used by the authors and others at Los Alamos for the study of different kinds of radiations.

Radioactive Measurements with Nuclear Emulsions

By Herman Yagoda. Pp. 356+ix, Figs. 75. John Wiley and Sons, Inc., New York, 1949. Price $5.00.

This volume is a description of the theory and practice of nuclear emulsion technique—the use of photographic emulsions in studying radioactivity. The author gives a detailed description of many practical methods for the use of such emulsions, and in addition he explains the theory which underlies the operations involved. The book describes carrier precipitation methods, emanations, alpha-star formation, and radiocolloid adsorption phenomena. The bibliography of 700 items is another valuable feature of the book.

Cosmic Ray Physics

By D. J. X. Montgomery. Pp. 370+viii. Princeton University Press, Princeton, 1949. Price $5.00.

This volume presents a survey of the entire field of cosmic-ray studies. Because of the uncertainty of theory in the regions of energy involved and the complexity of the phenomena, the emphasis of the book is experimental with theory subordinated. The various experimental techniques are discussed and explained, and numerous illustrations are used throughout. The book presupposes a knowledge of general physics equivalent to that of an upperclass undergraduate or a beginning graduate student.

On the Theory of Stochastic Processes and Their Application to the Theory of Cosmic Radiation

By Niels Arley. Pp. 240, Figs. 53. John Wiley and Sons, Inc., New York, 1949. Price $5.00.

The purpose of this book is to investigate in further detail the so-called fluctuation problem in the theory of cosmic radiation. Part I deals with the theory of stochastic, or random, processes. It is aimed at giving a further generalization of Feller's and Lundberg's theories for the purpose of application of stochastic processes to the theories of cosmic-ray showers. Part II discusses in detail the fluctuation problem in the theory of the cosmic-ray showers treated in the first part.

Introduction to Radiochemistry

By Gerhart Friedlander and Joseph W. Kennedy. Pp. 412+xiii. John Wiley and Sons, Inc., New York, 1949. Price $5.00.

Included in the early chapters of this book is descriptive information about atomic nuclei, nuclear reactions, and the machines used for the acceleration of nuclear projectiles. Later chapters introduce in logical sequence the quantitative treatment of radioactive decay processes, the various types of radioactive disintegration, the radiations emitted, their interactions with matter, and their detection. The book presupposes no previous knowledge of nuclear physics. At the end of each chapter are lists of references which call attention to standard works and recent literature on specific topics. Exercises are also included at the end of each chapter.

Radioactive Tracer Techniques

By George K. Schweitzer and Ira B. Whitney. Pp. 241+vi, Figs. 13. D. Van Nostrand Company, Inc., New York, 1949. Price $3.25.

This manual is designed to serve as a guide for laboratory work and instruction in the utilization of radioactive tracers. The first four chapters deal with topics pertaining to all radioactive laboratory work. Next basic experiments are presented. The remainder of the book is devoted to material illustrating chemical, physical, and biological tracer methods.

Isotopic Tracers and Nuclear Radiations with Applications to Biology and Medicine

By William E. Siri (with contributions by five others). Pp. 653+xiv, Figs. 136. McGraw-Hill Book Company, Inc., New York, 1949. Price $12.50.

This book has been developed to bridge the gap between those books intended solely for the nuclear physicist and those which merely describe the results of research in which radioactive isotopes and nuclear radiations were used. To this end, the authors have presented the theory, the methods of measurement, and the principles of applications of tracers and radiations without stressing unnecessary and complicated features of nuclear physics or rapidly outdated experiments in biology.

Fear, War, and the Bomb

By P. M. S. Blackett. Pp. 244. Whittlesey House, McGraw-Hill Book Company, Inc., New York, 1949. Price $3.50.

The author, winner of a 1948 Nobel prize, asserts that the world today has an atom neurosis. He believes that we are handcuffed by fear, and that before we can achieve lasting peace, that fear must be replaced by a firm belief in the long life of civilization. The author believes that the atomic bomb cannot bring about a cheap, quick military knockout in a war between Russia and the United States, nor is that war imminent. When this book was first published in Great Britain it aroused an immediate storm of controversy. The author, a former member of Great Britain's advisory committee on atomic energy, holds a position sharply at variance with the official attitudes of the British and American governments.

Constructive Uses of Atomic Energy

S. C. Rothmann, editor. Pp. 258+ix, Figs. 11. Harper and Brothers, New York, 1949. Price $3.00.

This book describes atomic energy applications in such fields as industrial power, chemistry, metallurgy, aviation, ceramics, soil-fertilizer research, and biological, pharmaceutical, and medical research. It is a collection of fourteen articles by scientists in atomic research.

SOLID STATE PHYSICS

Progress in Metal Physics, I

Bruce Chalmers, editor. Pp. 401+viii. Interscience Publishers, Inc., New York, 1949. Price $9.50.

The purpose of this volume, which is to be the first of an annual series, is to present authoritative reviews of the present state of knowledge in specialized aspects of the field that includes both physical metallurgy and metal physics. Nine authors have contributed chapters.

Fourier Technique in X-Ray Organic Structure Analysis

By A. D. Booth. Pp. 106. Cambridge University Press, London; The Macmillan Company, New York, 1948. Price $2.75.

This book gives a comprehensive account of the technique of organic crystal structure determination by Fourier methods, and

indicates the relative scope of the various techniques. The treatment is from fundamentals, but the reader is assumed to have an adequate knowledge of geometric crystallography and of the theory of space groups.

Crystals and X-Rays

By K. Lonsdale. Pp. 199+viii, Figs. 138. D. Van Nostrand Company, Inc., New York, 1949. Price $3.75.

This book is based on a course of public lectures given at University College, London, in the spring of 1946, and is designed to interest those who do not now use x-ray crystallography but who might well do so; and to instruct those who are making use of x-ray crystallographic methods.

On the Systems Formed by Points Regularly Distributed on a Plane or in Space

By M. A. Bravais. English translation by Amos J. Shaler. Pp. 113+viii, Figs. 42. The Crystallographic Society of America, care of Philips Laboratories, Irvington, New York, 1949. Price $3.40 to members, $3.90 to non-members.

With this volume the Crystallographic Society of America begins publication of its Memoir Series. This work by Bravais, concerned with the development of the fourteen space lattices, marked an important milestone in the science of crystallography.

The Adsorption of Gases on Solids

By A. R. Miller. Pp. 133+ix, Figs. 38. Cambridge University Press, New York, 1949. Price $2,50.

This book has been based upon "Some Problems in Adsorption" by J. K. Roberts, a *Cambridge Physical Tract* published in 1939. Since the original tract was written, certain statistical theories have been developed in connection with this work and other advances have been made. Hence it seemed desirable to rewrite the book. Nearly half of the present work is new material.

Prüfen und Messen der Oberflächengestalt

By Johannes Perthen. Pp. 257, Figs. 115. Carl Hanser-Verlag-München, Munich, 1949.

Metals Reference Book

Colin J. Smithells, editor. Pp. 735+xvi. Interscience Publishers, Inc., New York, 1949. Price $13.50.

The object of this book is to provide a convenient summary of data relating to metallurgy and metal physics. So far as possible the data are presented in the form of tables or diagrams with the minimum of descriptive matter, although short monographs have occasionally been included where information could not otherwise be adequately presented. The values given are those which contributors have selected as the most reliable after a critical review of the published data. A limited bibliography at the end of each section enables the reader to refer to the more important original sources.

Treatise on Powder Metallurgy, Vol. I. Technology of Metal Powders and Their Products

By Claus G. Goetzel. Pp. 778+xxvii, Figs 300. Interscience Publishers, Inc., New York, 1949. Price $15.00.

This is the first of three volumes intended to serve as an introduction into the field for the novice who wants detailed information, as a reference book for the initiated, and for the engineer who has to solve immediate problems.

Hydrogen in Metals

By Donald P. Smith. Pp. 367+xiii. The University of Chicago Press, Chicago, 1948. Price $10.00.

This monograph takes up the absorption and retention of hydrogen by massive metal. Rates of absorption, from gas and at the cathode, are discussed, as well as the factors which influence them. Existing solubility data are presented in numerous tables and graphs, with critical comment.

Metallic Creep and Creep Resistant Alloys

By A. H. Sully. Pp. 278+xii, Figs. 138. Interscience Publishers, Inc., New York, 1949. Price $5.50.

In this book the author has attempted to present the viewpoints of both the industrial worker and the academic researcher in the field of creep resistant alloys. The development of the existing theory of the strength of metals is described with special reference to creep, and an attempt has been made to explain certain metallurgical factors affecting creep on the basis of this theory. A detailed summary of experimental techniques for the study of creep phenomena has also been included as well as a survey of the creep properties of most of the well-known ferrous and non-ferrous alloys. Some speculations on future developments in creep resisting alloys are made.

Technology of Light Metals

By A. von Zeerleder. Pp. 366+xiv, Figs. 396. Elsevier Publishing Company, Inc., New York, 1949. Price $7.50.

This handbook deals with the working up of aluminum and magnesium alloys from the ore to the semimanufactured goods, taking into consideration the possibilities of the latest metallurgical improvements. To prevent this book from becoming too voluminous, some less important portions of the 1936 edition have been eliminated, with more space devoted to the most recent testing and working methods.

Metal Rectifiers

By H. K. Henisch. Pp. 156+xi, Figs. 55. Oxford University Press, New York, 1949. Price $3.75.

This is the second volume of a series intended to summarize the recent results of long-range research in materials and allied subjects, in a form useful to physicists in universities and industrial laboratories. The book deals with the theory and practice of dry rectification, especially those aspects of the subject which are parts of physics rather than of engineering. Attention is paid especially to the theories of rectification put forward by Schottky, Wilson, and others, and to the experimental work which is relevant to their verification.

Titanium

By Jelks Barksdale. Pp. 591+xii, Figs. 15. The Ronald Press Company, New York, 1949. Price $10.00.

The purpose of this book is to meet the need for a working reference volume on titanium. Abstracts of the literature on titanium were studied by the author, and essential information incorporated into the text. Exact references to all the material consulted are included.

Industrial Rheology and Rheological Structures

By Henry Green. Pp. 311+xii, Figs. 78. John Wiley and Sons, Inc., New York, 1949. Price $5.50.

Believed to be the first of its kind, this book is a manual and practical working guide for making rheological and structural microscopic investigations. It can be used by university or industrial laboratory where researchers deal with such articles as printing inks, paints, oils, varnishes, biochemicals, resins, plastics, food, cement, and asphalt.

Engineering Laminates

ALBERT G. H. DIETZ, editor. Pp. 797. John Wiley and Sons, Inc., New York, 1949. Price $10.00.

This book contains valuable information for everyone concerned with the design, manufacture, and use of structural materials. The book completely explains the mechanics of laminate materials from basic properties, through physical and chemical properties, to industrial uses. Discussions of wood, metallic, and rubber-based laminates are included. Throughout the book the authors have emphasized engineering properties, design, manufacture, and use. The book has a contributing staff of 24 engineers, architects, builders, metallurgists, designers, and professors.

High-Polymer Physics—a Symposium

HOWARD A. ROBINSON, editor. Pp. 572+xiii. Remsen Press Division, Chemical Publishing Company, Inc., Brooklyn, New York, 1948. Price $12.00.

Thirty-nine physicists and chemists, all rather prominent experts in their special fields, contributed to this symposim on hte relatively new subject of high-polymer physics. The results discussed come largely from industrial research laboratories rather than from college research groups. The physicist had to help the chemist, and vice versa, to explain the phenomena encountered in high polymers. This volume is the product of their cooperation.

APPLIED PHYSICS NOT INCLUDED ELSEWHERE

Mathematical Biophysics

BY NICOLAS RASHEVSKY (revised edition). Pp. 669+xxiii. University of Chicago Press, Chicago, 1948. Price $7.50.

This study is intended to lay the foundation for a mathematical biology, analogous in its methods to mathematical physics and based on the physical interpretations of biological phenomena. In this new edition the author has included recent developments in the field of mathematical biophysics and has incorporated most of the material formerly published in his *Advances and Applications of Mathematical Biology*. This edition contains 200 pages of new material, including the theory of embryonic development, the theory of cancer, and an elaboration on the theory of cell division. Also included are a number of new chapters on learning, abstractions, logical thinking, and perception of visual patterns.

Biological Reactions Caused by Electric Currents and by X-Rays

BY J. TH. VAN DER WERFF. Pp. 203, Figs. 38. Elsevier Publishing Company, Inc., New York, 1948. Price $5.00.

This book is a theoretical study of the phenomena of excitation in the nerve by different electric currents and of the biological reactions caused by x-rays, both based upon a common principle.

The Atmospheres of the Earth and Planets

GERARD P. KUIPER, editor. Pp. 366+vii, Figs. 91. The University of Chicago Press, Chicago, 1949. Price $7.50.

Physical Principles of Oil Production

BY MORRIS MUSKAT. Pp. 922+xv. McGraw-Hill Book Company, Inc., New York, 1949. Price $15.00.

The purpose of this book is to formulate and correlate present information concerning the physical principles and facts underlying the mechanics of oil production. A further purpose has been to stimulate and encourage additional research and study of the subject to fill in the many gaps in our present knowledge, clarify the many aspects which are still subject to speculation and conjecture, generalize the simplified and idealized treatments of special problems, and to improve the correlation between laboratory theories and field observations. The approach of the book consists in the detailed analysis of the physical concepts and principles which are developed as foundations for the interpretation and study of actual oil field performance. Many examples from actual oil field operation demonstrate the basic characteristics of the various types of producing mechanism and operation.

Seismicity of the Earth and Associated Phenomena

BY B. GUTENBERG AND C. F. RICHTER. Pp. 273+vii, Figs. 34. Princeton University Press, Princeton, 1949. Price $10.00.

This is an account of the relative seismicity of various parts of the earth during the limited period (since about 1904) for which sufficiently accurate data are available. Geographical and geological relations of the principal earthquake zones and areas are discussed, including correlation with mountain structures, oceanic deeps, gravity anomalies, and active volcanoes. A new checklist of active volcanoes is included. The origin of earthquakes is discussed, particularly with reference to folding and thrusting in the structural arcs of Pacific type.

Oscillations of the Earth's Atmosphere

BY M. V. WILKES. Pp. 76+ix, Figs. 23. Cambridge University Press, New York, 1949. Price $2.50.

This monograph discusses all that is known to date of the oscillations of the atmosphere. It treats of the lunar and solar air tides, their measurement, manifestations, and causes; the theory of oscillations in a rotating atmosphere; the numerical evaluation of air tides; the outward flux of energy; the height of the earth's atmosphere; the temperature variation, etc. Observational and mathematical methods of arriving at a knowledge of the properties of the atmosphere are referred to as they arise, and the book concludes with a general discussion of results.

Some Recent Researches in Solar Physics

BY F. HOYLE. Pp. 134+xii, Figs. 8. Cambridge University Press, New York, 1949. Price $3.00.

The aim of this monograph is to provide an account of the present state of solar physics. The first two chapters are largely devoted to an account of the observational material concerning sunspots, the solar cycle, the chromosphere, and the corona. In the third and fourth chapters a new theory of the origin and structure of the chromosphere and corona is presented, while in the remainder of the book such diverse topics as electromagnetism in the solar atmosphere, solar and terrestrial relationships, the emission of radio waves by the sun, and the origin of the general solar magnetic field are discussed.

The Solar System Analysed

BY F. C. ATTWOOD. Pp. 88. The Dawson Printing Company, Ltd., Auckland, New Zealand, 1947.

This brief book is subtitled "Old Laws from a New Angle." Chapters deal with the beginning of the solar system, the solar nebula, the origin of the earth's rotation, and so on through a concluding chapter named "Speculation."

Our Sun

BY DONALD H. MENZEL. Pp. 326, Figs. 178. The Blakiston Company, Philadelphia, 1949. Price $4.50.

In language that is easily understood, this volume explains the basic methods for finding out about the sun. It vividly pictures solar activity and the vast explosions occurring in the sun's atmosphere. Included are the many new advances in solar research heretofore restricted from publication because of military classification. The author discusses the relationship between the sun and our natural resources; its probability as a source of heat and fuel; the sun and atomic power; and its effect on radio transmission and weather forecasting.

The Face of the Moon

By Ralph B. Baldwin. Pp. 239+xiv, Figs. 26, Plates XVII. The University of Chicago Press, Chicago, 1949. Price $5.00.

"How and when did the moon arrive at its present state of dilapidated grandeur?" In this interesting volume the author examines and rejects all but one of the processes previously suggested to account for the lunar craters, rays, mountain ranges, and lava flows. He selects as the most probable cause the impacts of giant meteorites on the rocky lunar face and their resultant explosions. Chapters are devoted to a description of the strange and beautiful formations found on the moon, to a discussion of modern and ancient meteoritic craters known to exist on earth, to an analysis of the impact theory in the light of knowledge of man-made explosion pits, to an investigation of the possible effects of a lunar atmosphere, to the dating of the major changes on the face of the moon, and to a consideration of the possibility of meteoritic infalls on other planets.

Hydrology

By C. O. Wisler and E. F. Brater. Pp. 419+xi, Figs. 132. John Wiley and Sons, Inc., New York, 1949. Price $6.00.

This textbook presents the fundamental principles of hydrology as they have been developed up to the time of writing. The distribution graph, the unit hydrograph, the theory of infiltration capacity, and many other recent developments are presented. The many applications cited throughout the book make it useful to the general practicing engineer, forester, and agriculturist, as well as to the student who is looking for the basic principles of hydrology. The book provides information which can be used in the determination of the maximum flood flow that may be expected every few years, as in airport drainage and storm sewer design. For the design of a spillway for an important dam, or any flood control work, the authors present methods for determining the maximum flow that will occur with rare frequency, such as once in a thousand years.

Surface Tension and the Spreading of Liquids

By R. S. Burdon. Pp. 92+xi, Figs. 22. Cambridge University Press, New York, 1949. Price $2.50.

The first edition of this book was published in 1940. In this new edition the author has added accounts of all important work published recently. Chapter subjects include the nature of surface forces; measurement of surface tension; the surface of liquid metals; spreading; general conditions; spreading on the surface of mercury; spreading on water; liquids on the surface of solids.

Experimental Immunochemistry

By Elvin A. Kabat and Manfred M. Mayer. Pp. 567+xv, Figs. 88. Charles C. Thomas, Publisher, Springfield, Illinois, 1948.

The authors of this book have undertaken the task of assembling for the first time in one place the scattered techniques which have been developed to fill the needs of immunochemical problems. Main emphasis is placed on quantitative analytical methods. To meet the needs of students as well as of chemists who may not be thoroughly familiar with immunology, and of bacteriologists and immunologists who may not be well acquainted with microchemical methods, introductory and background material in both fields has been included.

Bibliography of Research on Heavy Hydrogen Compounds

Harold C. Urey and Isidor Kirshenbaum, editors; compiled by Alice H. Kimball. Pp. 350+xvi. McGraw-Hill Book Company, Inc., New York, 1949. Price $3.25.

This volume is a collection of about 2000 references to published literature on the subject of the heavy hydrogen isotope. The material is classified according to subject in such a manner that previous work in the field may be readily found. The volume should prove useful to anyone concerned with research on isotopes and their application to specific problems.

Physical Methods of Organic Chemistry, Parts I and II

Arnold Weissberger, editor (second edition). Pp. 1072+xii and 1024+xi. Interscience Publishers, Inc., New York, 1949. Price $12.50 each.

This familiar work has been completely revised and augmented, and as before, appears in two volumes. The new edition has new chapters on temperature measurement, temperature control, vapor pressure, the ultra-centrifuge, electrophoresis, and turbidimetry. The chapters on viscosity, osmotic pressure, diffusivity, dipole moments, radioactivity, and mass spectrometry have been entirely rewritten. All other chapters have been thoroughly revised.

Outlines of Physical Chemistry

By Farrington Daniels. Pp. 713+viii, Figs. 164. John Wiley and Sons, Inc., New York, 1948. Price $5.00.

This is the seventh revision of what was originally the well-known "Getman and Daniels" text on physical chemistry. The illustrations in this edition have been revised in keeping with the author's emphasis on practicality. Hypothetical diagrams illustrating physical chemical phenomena have been replaced with data from laboratory measurements. One-fourth of the numreous problems have been replaced or revised.

Sources of Engineering Information

By Blanche H. Dalton. Pp. 109. University of California Press, Berkeley, 1948. Price $4.00.

This is a guide to engineering literature and data designed as a concise reference for engineering students, practicing engineers, research workers, and librarians. Its arrangement is based on studies of the use of an actual library collection. The book includes information on every type of engineering.

Rocket Propulsion Elements

By George P. Sutton. Pp. 294+ix. John Wiley and Sons, Inc., New York, 1949. Price $4.50.

Airplane Performance, Stability, and Control

By Courtland D. Perkins and Robert E. Hage. Pp. 493+vii. John Wiley and Sons, Inc., New York, 1949. Price $7.00.

This book is divided into two major parts, the first dealing with the problems of airplane design for performance, and the second involving the design of the airplane for adequate stability and control characteristics. In the first part, basic aerodynamics is summarized so that an estimation of airplane drag can be made for subsonic, transonic, and supersonic speeds; propulsion characteristics of the reciprocating engine, turbine propeller, turbojet, ramjet, and rocket are presented; performance methods are analyzed for the basic forces acting on an airplane utilizing any of the above power plants. In the second part of the book, the comparatively new art of designing the airplane for adequate flying qualities has been dealt with, with only subsonic flight being discussed. A modern analysis of the dynamics of the airplane is included.

GENERAL AND HISTORICAL

Reports on Progress in Physics, Vol. XII

Pp. 382. The Physical Society, London, 1949. Price to non-Fellows £2 2s. (postage 1s.).

This volume, containing papers by 19 authors, includes the following subjects: mass spectrometry, nuclear paramagnetism,

phosphors and phosphorescence, spontaneous fluctuations, recent nuclear experiments with high voltage x-rays, linear accelerators, viscosity and related properties in glass, theory of the oxidation of metals, fracture and strength of solids, multipole radiation in atomic spectra, collisions between atoms and molecules at ordinary temperatures, low temperature physics, slow neutron absorption cross sections of the elements, and molecular distribution and equation of state of gases.

Fundamentals of Physical Science

By Konrad Bates Krauskopf (second edition). Pp. 676+xii, Figs. 330. McGraw-Hill Book Company, Inc., New York, 1948. Price $4.50.

The preparation of this second edition was prompted chiefly "by a desire to bring the book up to date in several rapidly advancing fields of physical science." Notable changes are a complete rewriting of the chapter on the atomic nucleus, addition of a section on the uncertainty principle, introduction of Brönsted's theory in the discussion of acids and bases, and increased emphasis on air-mass analysis in weather forecasting.

Physics in the Modern World

By Henry Semat. Pp. 434. Rinehart and Company, New York, 1949. Price $5.00.

This book is addressed specifically to students of the arts and social sciences who are interested in learning and understanding the methods of the physical sciences. This volume is the outgrowth of three years of experience in teaching "Physics in the Modern World" to non-science students at the College of the City of New York. (Similar courses of the same length are given in chemistry, biology, and geology-astronomy.)

The Nature of Physical Theory

By P. W. Bridgman (reprint of 1936 edition). Pp.138+vi. Dover Publications, New York, 1949. $2.25.

This book is a study of modern scientific theory which views the problem of correlating reality with the various theoretical structures set up in the past fifty years. It analyzes the inadequacies of physical theories in supplying the answers to all questions, but points out the real service these concepts render within limited frames of reference. The author investigates and weighs the contributions of Newton, Bohr, Einstein, Heisenberg, and others. He discusses such basic concepts as the wave and corpuscular theories, the "clocks" of relativity, the "laws" of probability, and the mathematical system itself.

Science at War

By J. G. Crowther and R. Whiddington. Pp. 185, Figs. 51, Plates 51. Philosophical Library, Inc., New York, 1948. Price $6.00.

This is a detailed account of science's contribution to the war effort, based on the official archives and documents assembled by the scientific advisory committee to the British Cabinet. This volume should be of fundamental interest to the American scientist as well as to the intelligent layman interested in science. Among the basic topics dealt with in the book are radar, the atomic bomb, operational research, science, and the sea.

The Principles of Scientific Research

By Paul Freedman. Pp. 222, Figs. 11. Macdonald and Company, Ltd., New York, 1949. Price $3.05.

This book is addressed chiefly to young men and women embarking on or concerned with scientific research, but it is also intended for business executives in whose power, to a considerable extent, the successful conduct of scientific research lies.

The Universe and Dr. Einstein

By Lincoln Barnett. Pp. 127. William Sloane Associates, New York, 1948. Price $2.50.

This book is a report on what the men at the forefront of science know about the nature of the universe. The story of the search for essential truth about the universe is here told simply, in the vocabulary of the intelligent layman. This book can be understood by the high school student of physics and chemistry and still could be read with profit by outstanding scientists of today.

Man and This Mysterious Universe

By Brynjolf Bjorset. Pp. 174+vii. Philosophical Library, New York, 1949. Price $3.75.

This small book is intended as a synthesis of the many aspects of modern civilization, bringing within its scope the contribution of the East as well as of the West, showing how it has grown from the civilization of the past and how it will probably develop in the future. The author has traveled for thirty years and lived in more than fifty countries.

After Dinner Science

By Kenneth M. Swezey. Pp. 182. Whittlesey House, McGraw-Hill Book Company, Inc., New York, 1948. Price $3.00.

For everyone interested in popular science, here is a book full of pictures and fascinating scientific experiments that can be performed with ordinary household materials and without any previous scientific knowledge. The author takes the mystery out of science. With photographs and easy-to-understand explanations he shows the reader how he himself can demonstrate the basic laws of scientific phenomena. This is a volume that the whole family can enjoy.

Mendeleyev—the Story of a Great Scientist

By Daniel Q. Posin. Pp. 345+xii. McGraw-Hill Book Company, Inc., New York, 1948. Price $4.50.

This is the first biography in English of this famous Russian chemist and physicist. It reveals an almost unbelievable story of triumph in a man with a burning desire to be a scientist. It was Mendeleyev who first unlocked the doors of Russia's great natural resources. As a young man he had begun his studies handicapped by illness and poverty. Studying and teaching in the daytime, he spent most of the night in his laboratory. He discovered a plan of unity among the chemical elements which led him to fromulate the Periodic Table. When he died at 73, he had been acclaimed by every country in the world.

Albert Einstein, Philosopher-Scientist

Paul Arthur Schilpp, editor. Pp. 781+xvi. The Library of Living Philosophers, Northwestern University, Evanston, Illinois, 1949. Price $8.50.

This volume contains Albert Einstein's autobiography. In addition, contributions from 25 scholars and scientists in eleven countries (among them six Nobel Prize winners) discuss various aspects of Einstein's scientific work and achievements. In Part III of the volume Einstein replies to the comments, analyses, and criticisms of his contemporaries. An exhaustive bibliography of Einstein's writings is included.

Scientific Autobiography and Other Papers

By Max Planck. Pp. 192. Philosophical Library, New York, 1949. Price $3.75.

This is a collection of the last writings of one of the greatest physicists of our time. The essays that make up this small volume deal with problems such as: true and fictitious issues in science; meaning and limit of exact science; physics and causality; natural sciences and the religious world. The autobiographical notes tell the story of Max Planck's own personal and scientific development.

Journal
of
Applied Physics

Volume 21, Number 5 May, 1950

On the Shielding Qualities of Different Concrete Mixtures*

P. C. Gugelot and M. G. White
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey
(Received April 4, 1949)

Absorption curves for neutrons and γ-radiation produced by 16-Mev protons on beryllium are measured in absorber blocks of various concrete mixtures to obtain information on their shielding properties. It is shown that a mixture of cement, scrap iron, and limonite ore is as efficient as water in slowing down and capturing neutrons and, in addition, has sufficient density ($\rho=3.5$ g/cm^3) to be an effective absorber for γ-radiation. Compared to ordinary concrete, a three-foot wall of iron and limonite concrete is superior by a factor of 280 for neutrons and by a factor of 20 for γ-radiation. Boron is shown to be of little value in this type of shielding. Mechanical properties of the scrap iron-limonite concrete are satisfactory for structural purposes. From the practical cost standpoint, a concrete made entirely of limonite ore is shown to be a reasonable compromise.

I. INTRODUCTION

With the construction of larger accelerators and nuclear reactors, the problem of shielding against radiation becomes more and more important. It is not only necessary to protect against health hazards, but also to reduce the background for experiments.

Water shields commonly used around cyclotrons are effective in slowing down and capturing neutrons of a few Mev. However, for very high neutron energies the slowing down qualities of this medium decreases, owing to the small scattering cross section of fast neutrons in hydrogen. In addition, the absorption of γ-radiation in water is negligible. It is, therefore, necessary to find a shielding material which is effective in slowing down neutrons, and with sufficient density to be an efficient γ-ray absorber.

The theory of slowing down of neutrons has been worked out in a hydrogenous medium by Fermi.[1] His theory describes the mean square distance $(r^2)_{Av}$ which a neutron with the initial energy E_0 travels in being slowed down to an energy E_1. The "Age-theory," developed by Bethe, Korf, and Placzek,[2,3] calculates the spatial distribution of neutrons in a medium. By using the age of a neutron:

$$\tau=1/6(r^2)_{Av}=\frac{1}{3[1-(2/3A)]\xi}\int_{E_1}^{E_0}\lambda^2\frac{dE}{E}$$

$$\xi=1-\frac{A}{2}[1-(1/A)]^2\log(A+1)/(A-1),$$

where A = atomic weight of element of which the absorber is composed, and λ = mean free path of the neutrons of energy E, the slowing down equation can be written in a simple form:

$$\Delta q=\partial q/\partial\tau.$$

q = number of neutrons per cm^3 slowed down to energy E_1, the solutions of which are Gaussian functions. Placzek[4] and Marshak[5] extended the theory to absorbers composed of heavy nuclei, and to mixtures of different nuclei. Higher approximations were developed showing that the Gaussian distribution of neutrons of final energy E_1, obtained by the age-theory, is valid only to distances equal to τ/λ and then only when many collisions are necessary to slow the neutrons down ($A\gg1$). At larger absorber thickness the neutron density decreases exponentially. These calculations are all based on slowing down of neutrons by elastic collisions;

* Assisted by the joint program of the ONR and the AEC.
[1] E. Fermi, Ricerca Scient. 7, 13 (1936); H. A. Bethe, Rev. Mod. Phys. 9, 69 (1937).
[2] Bethe, Korf, and Placzek, Phys. Rev. 57, 573 (1940).
[3] E. Fermi, Los Alamos Lectures (February, 1946).
[4] G. Placzek, Phys. Rev. 69, 423 (1946).
[5] R. E. Marshak, Rev. Mod. Phys. 19, 185 (1947).

therefore, they are only valid for initial neutron energies lower than the first excited states in nuclei. The poor knowledge about nuclear excitation levels makes calculations above these energies uncertain and experiments are needed.

Measurements were carried out by Munn and Pontecorvo[6] with heterogeneous mixtures of water with iron or bismuth. They obtained the mean square distance $(r^2)_{Av}$ required to slow Ra+Be neutrons down to 1.44 ev (In detectors) and to thermal energy (Dy detectors). Values observed by these authors are:

for water: $(r^2)_{Av\text{-}In}=278$ cm^2, $(r^2)_{Av\text{-}th}=330$ cm^2,
for water+Bi (density 5.3 g/cm^3): $(r^2)_{Av\text{-}th}=562$ cm^2,
for water+Fe (density 3.9 g/cm^3): $(r^2)_{Av\text{-}In}=319$ cm^2,

showing that it is possible to use mixtures of water and iron in which $(r^2)_{Av}$ is nearly the same as in pure water, but with the greatly improved γ-ray absorption of the iron.

Tittle and Goodman,[7] in order to carry out absorption experiments in which the components of the mixture are easily interchangeable, studied the difference in neutron distribution between homogeneous and heterogeneous mixtures of C, O, and H. Neutrons from a Ra+Be source were used and In-resonance neutrons detected. The intensity distribution was found to be the same for the different mixtures with equal C, O, and H concentrations at distances >20 cm from the source.

Hildebrand and Moyer[8] obtained absorption coefficients for 80-Mev neutrons detected by an ionization chamber. The half-value thickness for water is 55.9 cm and for concrete (ordinary and barite) 24.1 cm.

Concrete with 5–10 percent water was also investigated by Jensen and Ritter[9] with neutrons from a Ra+Be source. Spherical and plane geometry were used. In the first case, a distribution for In-resonance neutrons was obtained of the form $r^2I=$ const. $\exp(-r/L')$ ($r>40$ cm), $L'=21.8$ cm. For the plane geometry the intensity decreased exponentially: $I=$ const. $\exp(-r/L)$ ($r>40$ cm), $L=15.3$ cm. The authors show that, at absorber thickness where the fast neutron flux is in equilibrium with the slow neutron flux for spherical geometry, $Ir=$ const. $\exp(-r/L_s)$, in which $L_s=L$, and $1/L'=1/L-1/r$. A water wall of 65-cm thickness reduces the neutron flux in this case as much as 160 cm concrete ($I/I_0=10^{-3}$).

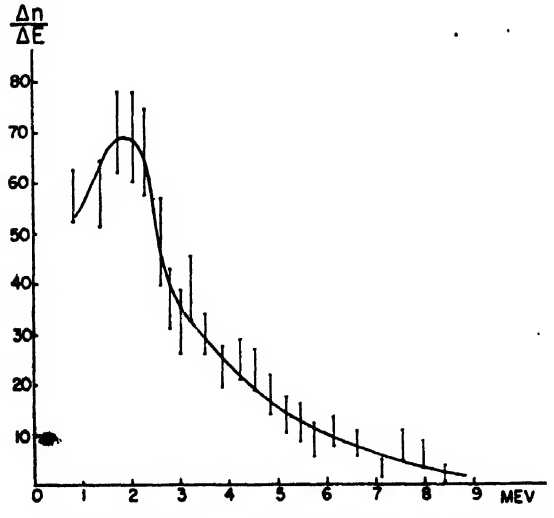


FIG. 1. Neutron spectrum from 16-Mev protons on a thick Be target.

II. EXPERIMENTAL ARRANGEMENT

Before the radiation shield around the Princeton frequency modulated cyclotron[10] was constructed, we made absorption measurements with different concrete mixtures in order to design a wall which would reduce background sufficiently to make possible careful experiments with the deflected proton beam. Cubic blocks of 46-cm edge were cast, having a steel tube of 7.5-cm diameter through the center of the block. Slabs of 5.3-cm thickness, prepared from the same mixture, fitted smoothly into the tube. In this way slow neutron detectors or a Geiger-Müller counter could easily be placed between the slabs and an almost homogeneous absorber was obtained. The absorber blocks were placed in a reproducible position in the forward direction of neutrons generated by 16-Mev protons on a thick Be target. To obtain information on the energy of the neutrons and to be able to compare our results with the experiments which were carried out with a Ra+Be neutron source, we studied the neutron spectrum by measuring proton-recoil tracks in 100μ Ilford C2 photographic emulsions (Fig. 1). Tracks in the forward direction, originating and ending in the emulsion, were measured and corrected for scattering cross section[11] and for loss of tracks in a layer of finite thickness. The plate was bombarded outside the acceleration chamber 52 cm from the internal Be target. An appreciable number of the low energy tracks is due probably to neutrons which suffered collisions in the iron of the magnet. This spectrum differs from a Ra+Be neutron spectrum[12] in the low energy range; however, the high energy tail is in both cases somewhat similar. In a total of 900 tracks, five were found with energy >9 Mev.

To be effective, a shield must strongly absorb fast neutrons, slow neutrons and gamma-rays since all these particles are radiated by nuclear machines and also result from the stopping of fast neutrons. A complete

[6] A. M. Munn and B. Pontecorvo, Can. J. Research A25, 157 (1947).
[7] C. W. Tittle and C. Goodman, Phys. Rev. 73, 1264 (1948).
[8] R. Hildebrand and B. J. Moyer, Phys. Rev. 72, 1258 (1947).
[9] P. Jensen and O. Ritter, Zeits. f. Naturforschung 2a, 376 (1947).
[10] White, Fulbright, Gugelot, and Bush, Phys. Rev. 74, 1242A (1948).
[11] Goldsmith, Ipsen, and Feld, "Neutron cross section of the elements," Rev. Mod. Phys. 19, 259 (1947); D. Bohm and C. Richman, Phys. Rev. 71, 567 (1947).
[12] D. L. Hill, Argonne National Lab., Declassified Report No. CP-3800.

study of shielding materials should include attenuation measurements on all these types of incident radiation using detectors which are specific for the radiation in question. However, for neutron studies, we have confined ourselves to the use of thermal or silver resonance neutron detectors and for gamma-rays a copper walled Geiger-Müller counter was used. The attenuation of fast neutrons we infer from the slopes of the slow neutron intensity curves at large absorber thickness since all *incident* slow neutrons are removed by the early part of the absorber and only fast neutrons are left to produce the detected slow neutrons. Considerable care must be exercised in using data obtained near the back of a block because room-scattered low energy neutrons can falsify the curves.

The slow neutron intensity in the absorber blocks due to the total incident cyclotron flux was obtained by activating silver foils of 2.5-cm diameter and 0.7-mm thickness between the slabs and the intensity of the 2.37^m Ag^{108}-period was measured with a thin wall G-M tube. The counting rate was corrected for decay and background and then compared with the activity of a monitor foil. To obtain the activation by resonance neutrons, the Ag foils were placed in an envelope of 0.45 g/cm^2 Cd (Ag^{108} has resonance levels at 50 and 15 ev). Silver was chosen because of its convenient half-life. The size of the punched foils was identical. In general, only the front side of the foils was measured. All these data are given in the "*B*" curves of Figs. 3–10. Runs in which the activity on both sides was investigated showed that the error introduced by measuring only one side was less than 4 percent. An uncertainty in foil position of 1–2 mm resulted from the rather rough faces of the slabs.

The diffusion of *thermal* neutrons incident on the test blocks was studied by interposing water cans between the cyclotron and test blocks and then taking attenuation curves with and without cadmium absorber placed *between* the water cans and the test block. These data are the "*A*" curves in Figs. 3–10. According to the diffusion theory, for plane geometry the intensity of thermal neutrons decreases exponentially. The $1/e$-value thickness is the "diffusion-length" l; it is connected with the transport mean free path λ and capture mean free path Λ by $l=(\frac{1}{3}\lambda\Lambda)^{\frac{1}{2}}$.

The influence of the finite size of the test blocks on the absorption curves was estimated in the following way: The neutron attenuation curve obtained in the cement +iron+limonite ore test block was compared with the attenuation curve obtained by placing water cans in front of the test block leaving an opening of 15×15 cm in the center. This in effect decreases the size of the absorbing block. That this is the actual case is shown by the slow neutron intensity distribution perpendicular to the fast neutron beam (Fig. 2). The neutron density obtained with the water can "diaphragm" (curve 2) has a pronounced maximum in the center, and the slope of the attenuation curve was unchanged from that

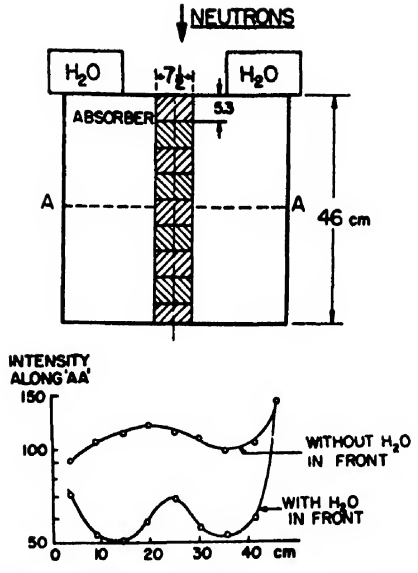


FIG. 2. Absorber block and thermal neutron distribution inside the absorber perpendicular to the fast neutron flux, showing the effect of neutron collimation by water tanks in front of the absorber.

obtained when the entire block was irradiated. The slow neutron intensity at large thickness of absorber is mainly caused by neutrons slowed down by a large number of small angle deflections, and the subsequent diffusion of slow neutrons contributes very little to the observed half-value thickness of the attenuation curve.

Gamma-ray attenuation, when measured by exposing test blocks to the entire cyclotron radiation and studying the counting rate of a copper walled Geiger-Müller counter cannot be expected to yield precise results since a few counts will be due to fast neutrons causing knock-on ions in the counter. More important, the fast neutrons will create gamma-rays in the test block. Nevertheless, as a practical matter, the attenuation of a shield as measured by a Geiger-Müller counter is of considerable interest. Figure 11 collects all the Geiger-Müller counter data for the various absorbers.

III. RESULTS FROM NEUTRON ABSORPTION EXPERIMENTS

a. Water

For this experiment a tank was used of larger dimensions than the concrete absorbers. The neutron detectors were placed in a hard rubber holder. Figure 3 shows the absorption curves so obtained. Figure 3A gives the attenuation of *thermal* neutrons which strike the water tank after being generated in a preceding layer of water. Curve 1 corresponds to the intensity distribution resulting from *all* incoming neutrons, curve 2 to the intensity distribution resulting from the faster than thermal neutrons which succeed in passing a cadmium sheet placed between the test water tank and the slow neutron generating water cans. The difference of these two curves shows the exponential decrease in

the intensity of incoming thermal neutrons. The diffusion-length is $l=2.7$ cm; it compares within the error to the value 2.88 cm obtained for water by Allison *et al.*[13] with In detectors. The deviation is caused by the uncertainty in position of the Ag-foils and by the Cd shield which eliminates neutrons reflected by the water cans back into the tank. Curve 3, of Fig. 3B, shows the total Ag activity caused by exposing the test tank to the direct cyclotron radiation, curve 4 the intensity produced by resonance neutrons only and the difference of 3 and 4 corresponds to the thermal neutron intensity. This curve yields a half-value thickness of 5.5 cm for absorber thickness larger than $r=20$ cm. From 3B it is possible to calculate $(r^2)_{Av}$ for thermal and resonance neutrons:

$$(r^2)_{Av}=\frac{\int_0^\infty I(r)r^2dr}{\int_0^\infty I(r)dr}$$

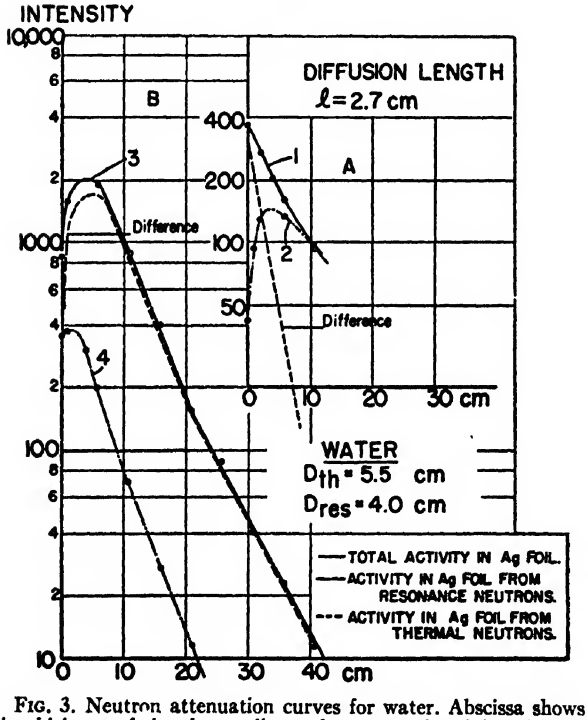


FIG. 3. Neutron attenuation curves for water. Abscissa shows the thickness of absorber, ordinate the corrected activity of silver detectors. A. Attenuation curves for incoming thermal neutrons, a layer of 15-cm water is placed between absorber and cyclotron. Curve 1. No Cd between absorber and waterlayer. All neutrons after passing through the 15-cm thick waterlayer can enter the absorber. Curve 2. Cd between absorber and waterlayer. Only faster than thermal neutrons enter the absorber tank. The difference of 1 and 2 shows the attenuation of incoming thermal neutrons from which the diffusion length of thermal neutrons is calculated. B. Attenuation curve of incoming fast neutrons. The absorber is only shielded by Cd. Curve 3. Ag detectors sensitive to all neutron energies. Curve 4. Ag detectors wrapped in Cd and thus are only sensitive to faster then thermal neutrons. The difference between 3 and 4 shows the Ag activity from thermal neutrons. From this curve the half-value thickness D_{th} is obtained.

[13] See J. H. Rush, Phys. Rev. 73, 271 (1948).

$I(r)$: The activity of the detectors, which is proportional to the slow neutron-intensity at position r.

For thermal neutrons: $(r^2)_{Av\text{-}th}=119.2$ cm².

For resonance neutrons: $(r^2)_{Av\text{-}res}=62.3$ cm².

For the Ag detector no precise neutron energy is connected with $(r^2)_{Av\text{-}res}$. The corresponding migration area is for thermal neutrons:

$$M_{th}^2=1/6(r^2)_{Av\text{-}th}=19.9 \text{ cm}^2.$$

The age for resonance neutrons is:

$$\tau_{res}=1/6(r^2)_{Av\text{-}res}=10.4 \text{ cm}^2.$$

According to the diffusion theory we have: $M_{th}^2=\tau_{th}+l^2$. The value for l is 2.88 cm. Hence the age of thermal neutrons is $\tau_{th}=11.8$ cm². The difference in age between thermal and resonance neutrons is: $\tau_{th}-\tau_{res}=1.3$ cm². This value is of the expected order and would correspond to an average resonance energy of about 30 ev. Comparing the age of our thermal neutrons with τ_{th} for thermal neutrons from a Ra+Be source slowed down in water, for which Anderson *et al.*[14] obtained 46.4 cm², shows that we have to deal with a neutron spectrum which extends to very low energies. It was indeed necessary to shield the front of the absorber with Cd to obtain useful values for $(r^2)_{Av\text{-}th}$.

b. Water+Boron

It has frequently been suggested that water shields could be improved by adding boron in the form of borax. Two beneficial effects are to be expected. First,

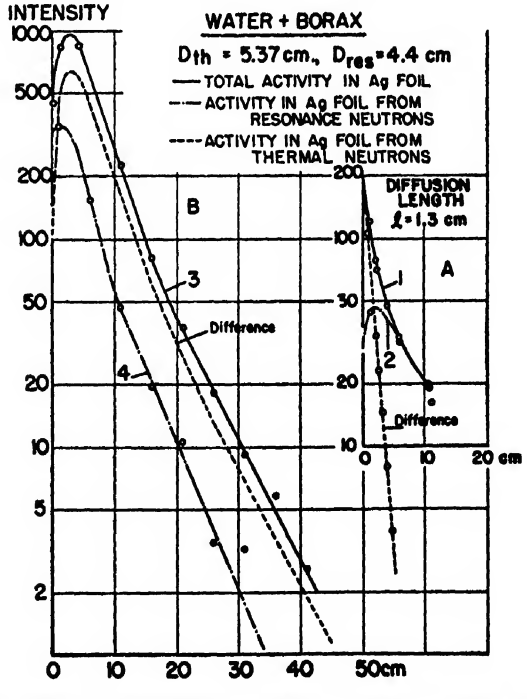


FIG. 4. Neutron attenuation curves for water+3 percent borax.

[14] See J. H. Rush, Phys. Rev. 73, 271 (1948).

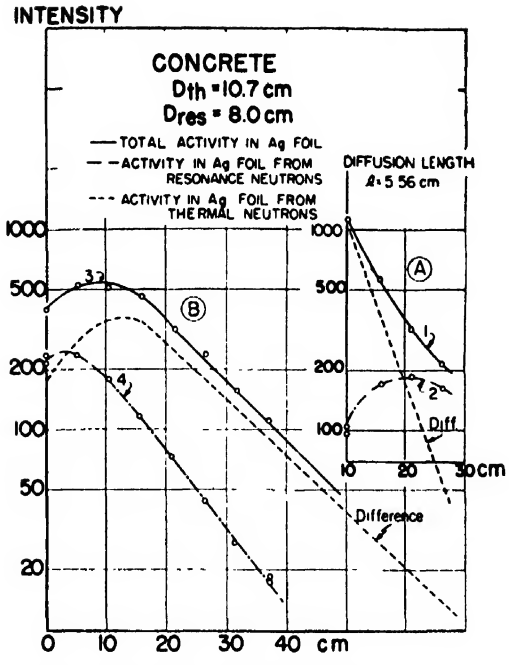


FIG. 5. Neutron attenuation curves for concrete.

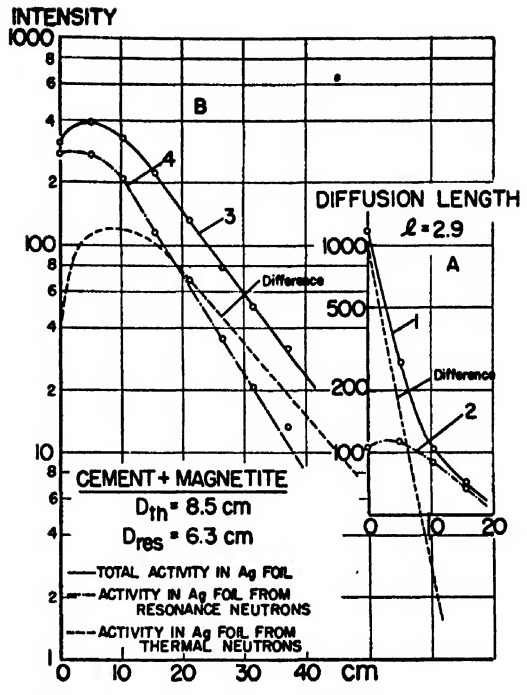


FIG. 6. Neutron attenuation curves for concrete+Pyrex.

the boron capture of neutrons tends to suppress the neutron-proton capture gamma-rays and second, the large boron cross section should reduce the diffusion of slow neutrons. To test the latter point a 3 percent solution of borax in water was substituted for the previously studied pure water. The capture mean free path Λ can be calculated assuming $\sigma_{H_a}=0.4\times10^{-24}$ cm^2 [15,16] and $\sigma_{B_a}=700\times10^{-24}$ cm^2. We obtain

$$\frac{\Lambda \text{ solution}}{\Lambda \text{ water}}=\frac{1}{6.2}.$$

From this it follows that

$$\frac{l \text{ solution}}{l \text{ water}}=\left(\frac{\Lambda \text{ solution}}{\Lambda \text{ water}}\right)^{\frac{1}{2}}=\frac{1}{2.5}.$$

By experiment (Fig. 4A) we observe l solution $=1.3$ and for pure water l water $=2.85$ (Fig. 3A) yielding the ratio

$$\left(\frac{l \text{ solution}}{l \text{ water}}\right)_{\text{exp}}=\frac{1}{2.2}.$$

The agreement between the predicted and observed ratio is reasonably good. However, the half-value thickness obtained from Fig. 4B, which includes *all* neutrons both fast and slow incident upon the borax solution absorber is the same as found for pure water showing that the slope of the attenuation curves is mainly determined by the larger mean free path of the fast neutrons present in the cyclotron radiation.

[15] Manley, Haworth, and Luebke, Phys. Rev. 61, 152 (1942).
[16] E. J. Harringh and J. L. Stewart, Can. J. Research 19, 33 (1941).

It is of interest to compare the effects of boron absorption on the flux of thermal neutrons and silver resonance neutrons. The area under an absorption curve, such as curve 3 Fig. 3B, is *proportional* to the integral

$$F_E=\int_0^\infty I_E(r)dr$$

where $I_E(r)$ is the number of neutrons at thickness r which have been slowed down to the energy E at which the detector is efficient. Thus if there is strong absorption at an energy *higher* than E the integral will be small. By extrapolating the curves of Figs. 3B and 4B we obtain for resonance energy neutrons the following data: In water $F_{res}=2.7\times10^3$; in water plus borax $F_{res}=2.3\times10^3$ showing that in spite of the boron the density of Ag *resonance* neutrons remain unchanged. However, for *thermal* neutrons $F_{th}=2.1\times10^4$ in water, and $F_{th}=5.6\times10^3$ in borax solution indicating that the boron has decreased the number of neutrons in the thermal energy range.

c. Concrete and Concrete+Pyrex

A sample of ordinary concrete was tested in the manner described above. The results are given in Fig. 5. The diffusion length for thermal neutrons is $l=5.6$ cm. The half-value thickness for the neutrons slowed down to thermal velocity is $D_{th}=10.7$ cm. This mixture contains 0.21 g water/cm^3, a number obtained from the amount of water used to make the mixture and from the moisture in the gravel and sand. (A small amount of

water will have evaporated during the setting of the concrete.) Composition: 46.3 percent gravel by weight, 27.6 percent sand, 17.7 percent cement and 8.4 percent water; density $\rho=2.35$ g/cm³. The large half-value thickness is due to the large number of collisions with relatively heavy nuclei which neutrons have to suffer in being slowed down; inelastic scattering will be negligible for Si and O.

Another sample was made containing 1.25 percent Pyrex glass by weight, 46.55 percent gravel, 27.78 percent sand, 18.37 percent cement and 6.05 percent water. The boron had an observable effect on the diffusion length (Fig. 6): $l=3.3$ cm (Pyrex contains 3 percent boron). The ratio of hydrogen atoms to boron atoms is in this case 2.7×10^{2}. If boron and hydrogen were solely responsible for capturing thermal neutrons the ratio of the diffusion lengths would be, taking into account the different hydrogen contents in both mixtures,

$$\frac{l\ \text{concrete+Pyrex}}{l\ \text{concrete}}=\frac{1}{2.7};$$

observed is

$$\left(\frac{l\ \text{concrete+Pyrex}}{l\ \text{concrete}}\right)_{\text{exp}}=\frac{1}{1.7}.$$

This deviation is to be expected because the thermal neutron capture cross sections of Si ($\sigma_a=0.25\times10^{-24}$ cm²) and of other elements of which the absorber is composed were not taken into account. The half-value thickness of the fast neutron attenuation curve decreased to $D_{th}=7.4$ cm. It is not clear whether the Pyrex had any function in decreasing the absorption coefficient D_{th}. This mixture contains only 0.14 g water/cm³. (Density $\rho=2.39$ g/cm³.)

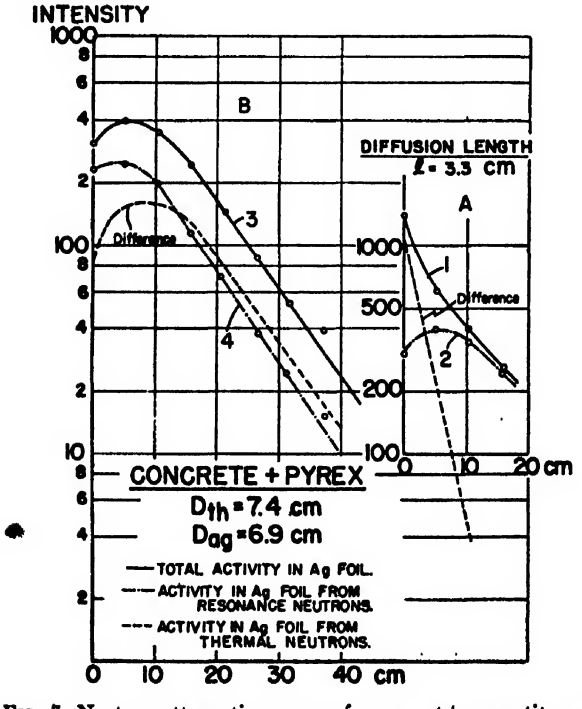


FIG. 7. Neutron attenuation curves for cement+magnetite ore.

d. Cement+Magnetite

No improvement was obtained by replacing the sand and gravel by magnetite ore (Fe_3O_4). This mixture contains 83.5 percent magnetite ore, 11 percent cement, and 5.5 percent water by weight. This amounts to 0.21 g water/cm³ and 1.93 g Fe/cm³; the density is 3.78 g/cm³. The half-value thickness is 8.5 cm (Fig. 7), slightly better than for standard concrete. Because of the iron used in this mixture the diffusion length for thermal neutrons became 2.9 cm. Fe has a relatively large capture cross section for thermal neutrons, $\sigma_{Fe_a}=2.05\times10^{-24}$ cm².[17]

e. Cement+Limonite, Cement+Limonite+Iron, Cement+Limonite+Iron+Pyrex

It was possible to increase the amount of hydrogen in these mixtures by using limonite ore ($2Fe_2O_3\cdot3H_2O$).*

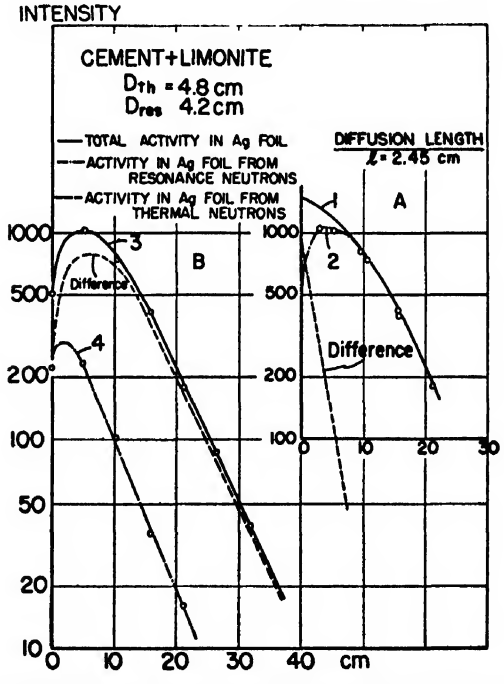


FIG. 8. Neutron attenuation curves for cement and limonite ore

The first absorber contains 68.5 percent limonite, 21 percent cement and 10.5 percent water. Taking the amount of moisture and water of crystallization into account 0.49 g water/cm³ is obtained and 0.93 g Fe/cm³. The density is $\rho=2.63$ g/cm³. The smallest half-value thickness D_{th} was obtained with this absorber: $D_{th}=4.8$

[17] J. W. Coltman and M. Goldhaber, Phys. Rev. 69, 411 (1946).

* Obtained from the Shook-Fletcher Company, Birmingham, Alabama.

cm (Fig. 8). However, the density was not sufficient for satisfactory γ-ray absorption (Section IV). Scrap iron in the form of punchings (about 2 cm diameter and about 1 cm thick) was mixed into the following block (Fig. 9): scrap iron 59.4 percent, limonite 26.35 percent, cement 9.54 percent and water 4.71 percent. A density of 4.41 g/cm^3 was obtained at the expense of lower hydrogen content: 0.35 g water/cm^3 and 3.23 g Fe/cm^3. The half-value thickness D_{th} increased to $D_{th}=5.4$ cm.

For the last block a mixture was obtained with less iron and more hydrogen than in the former absorber. The capture probability for thermal neutrons was again increased by the use of Pyrex: scrap iron 41.4 percent,

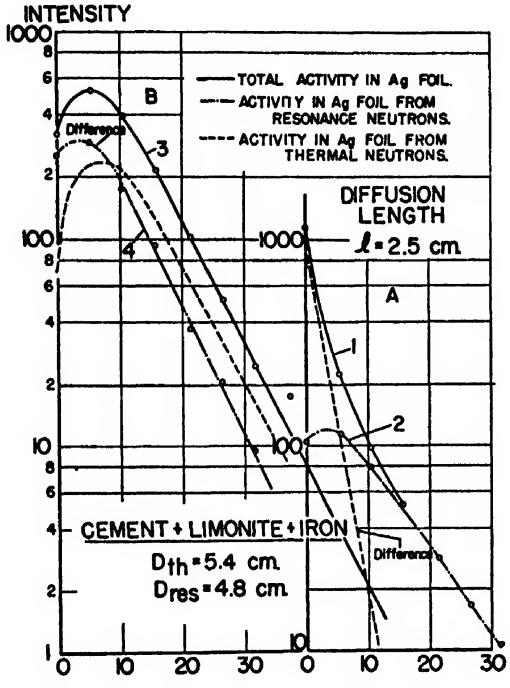


FIG. 9. Neutron attenuation curves for cement +limonite+scrap iron.

limonite 32.8 percent, cement 12.3 percent, Pyrex 0.7 percent, and water 12.8 percent. The density is 3.6 g/cm^3. This amounts to: 0.45 g water/cm^3 and 2.26 g Fe/cm^3. The half-value thickness obtained for thermal neutrons is $D_{th}=5.0$ cm. This smaller D_{th} is probably caused by the larger amount of water per cm^3. Because of the Pyrex used in this absorber, the diffusion length for thermal neutrons became $l=2.3$ cm, instead of 2.5 cm for the former two blocks (Fig. 10). In the cement +iron+limonite+Pyrex block the capture mean free path is, for hydrogen, 83 cm; for iron, 20 cm; for boron, 35 cm.

IV. γ-RADIATION

A Geiger Müller counter was placed between plugs in the test block to obtain information on the background which is to be expected outside the radiation shield when working with counters. A mica window "Victoreen" counter was used with the window closed by a brass cover. The counting rate is also connected with the neutron flux in the absorber because of (n, γ) and $(n, n\gamma)$ processes near the counter and (n, p) and (n, α) reactions in the counter wall. Therefore, with a bad neutron absorber, no steep absorption curves for the counter can be obtained. Figure 11 shows the results for all investigations. No simple exponential relation exists between intensity and absorber thickness. The largest half-value thickness, D_C, was observed for water. D_C remained unchanged in the boron solution, because only multiplication of the intensity by a constant factor <1 is to be expected for the neutron-capture γ-radiation. (All curves are normalized at $r=0$.) 1.25 cm of lead around the counter decreased the counting rate in water to 1/3.4 of its value at $r=0$ and to 1/4.1 at $r=30$ cm. More information is needed about the γ-ray spectrum and the fraction of counts caused by γ-radiation if one is to make detailed calculations.

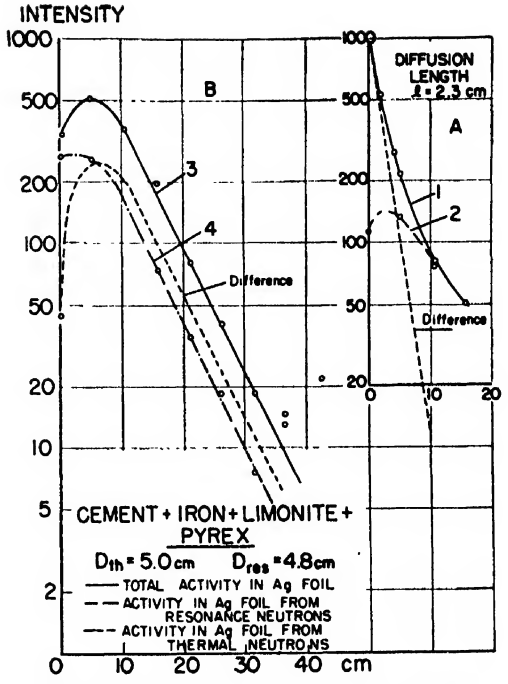


FIG. 10. Neutron attenuation curves for cement +limonite+scrap iron+Pyrex.

For both cement+limonite and cement+magnetite mixtures the same curves were observed. In the more dense magnetite the γ-ray absorption is probably higher than in the former, but this effect is obliterated by low neutron absorption in the magnetite block.

Almost exponential decays with half-value thickness of 7.9 cm are obtained in the limonite+iron absorbers.

V. DISCUSSION

By placing a water shield between an absorber block and the cyclotron, we noticed that the half-value thick-

TABLE I. Summarized data from absorption measurements.

Absorber	g H_2O/cm^3	g Fe/cm^3	Density g/cm^3	½ intensity thickness cm	Neutrons "Age" (cm^2) of thermal neutrons	Diffusion length (cm)	G-M counting rate ½ intensity thickness at $d=20$ cm	Transmission factor for a 90-cm thick wall: Neutrons	Transmission factor for a 90-cm thick wall: Counting rate
H_2O	1	—	1.00	5.5	11.8	2.7	14.8	$9\cdot10^{-4}$	$3\cdot10^{-2}$
H_2O+3% Borax	1	—	1.03	5.4	12.0	1.3	14.8	$6\cdot10^{-4}$	$3\cdot10^{-2}$
Concrete	0.2	—	2.35	10.7	86	5.6	13.3	$4.5\cdot10^{-3}$	$1\cdot10^{-2}$
Concrete+1.2% Pyrex	0.14	—	2.39	7.4	49	3.3	13.3	$3.8\cdot10^{-4}$	$7\cdot10^{-3}$
Cement+magnetite ore	0.21	1.93	3.78	8.5	70	2.9	9.0	$1\cdot10^{-4}$	$2\cdot10^{-3}$
Cement+limonite ore	0.49	0.93	2.63	4.8	23.7	2.5	9.0	$4.8\cdot10^{-4}$	$2\cdot10^{-3}$
Cement+Fe+limonite ore	0.35	3.23	4.41	5.4	32.9	2.5	7.9	$1.6\cdot10^{-5}$	$5\cdot10^{-4}$
Cement+Fe+lim+1% Pyrex	0.45	2.26	3.60	5.0	23.3	2.3	7.7	$6.4\cdot10^{-4}$	$4\cdot10^{-4}$

ness D_{th} increased. Water becomes increasingly transparent for the higher energy neutrons; so the addition of a water layer gives essentially the absorption coefficient for the high energy neutrons in the spectrum. For the cement+limonite+iron absorber (Fig. 9B) a value of $D_{th}=5.4$ cm was obtained without the water shield, but with 15 cm of water in *front* of the block $D_{th}=7$ cm at $r=20$ cm (Fig. 9A). This effect and the observed variations in the half-value thickness of the attenuation curve *in water* (at $r=10$ cm, $D_{th}=4.0$ cm and at $r=30$ cm, $D_{th}=5.5$ cm) showed that considerable care must be exercised in the interpretation of our data. It became necessary to investigate whether or not the half-value thickness of 5.4 cm in the limonite+iron mixture corresponded to the absorption of neutrons from the same part of the energy-spectrum as that which gave rise to a $D_{th}=5.5$ cm in water. The slope of the absorption curve in the limonite+iron block was investigated up to larger thickness of absorber by placing the cement+magnetite block in front of it. $D_{th}=5.4$ cm was obtained for the first 20 cm in the second absorber. At larger distances the intensity started to increase again; the neutron flux being so small that diffusion of almost thermal neutrons through the sides of the absorber became observable.

Thus, it cannot be excluded that the attenuation curve will become less steep with increasing absorber thickness, because of the slowing down process of a small fraction of very high energy neutrons. Insufficient knowledge about the various possible ways in which fast neutrons can be slowed down, e.g., inelastic scattering $(n, 2n)$ reactions, makes a detailed discussion about the form of an absorption curve impossible. However, the comparison of our different absorption curves is justified as long as can be shown that in each absorption experiment the same number of neutrons is slowed down to resonance energy. Thus, the number of neutrons slowed down to resonance energy gives evidence whether in some cases an appreciable number of fast neutrons is not sufficiently slowed down to E_{res} and thus fail to be detected. In Section III we found that

$$F_E=\int_0^\infty I_E(r)dr$$

is proportional to the total number of neutrons slowed down to the energy E. For all absorbers studied F_{res} amounts to *ca.* 3.10^3 (water: $F_{res}=2.73\times10^3$, for cement +iron+limonite+Pyrex; $F_{res}=3.51\times10^3$). This proves that in all cases about the same number of neutrons is slowed down to the silver *resonance* energy. The essentially constant value for F_{res} shows that capture of neutrons faster than E_{res} is in general negligible. The value F_{th} for thermal neutrons differs in some cases by a factor 3, because slow neutrons are absorbed before being detected.

A comparison with other experiments shows that absorption coefficients obtained with Ra+Be sources and those obtained with neutrons from Be+16-Mev protons are almost the same. By plotting $I\cdot r$ from the

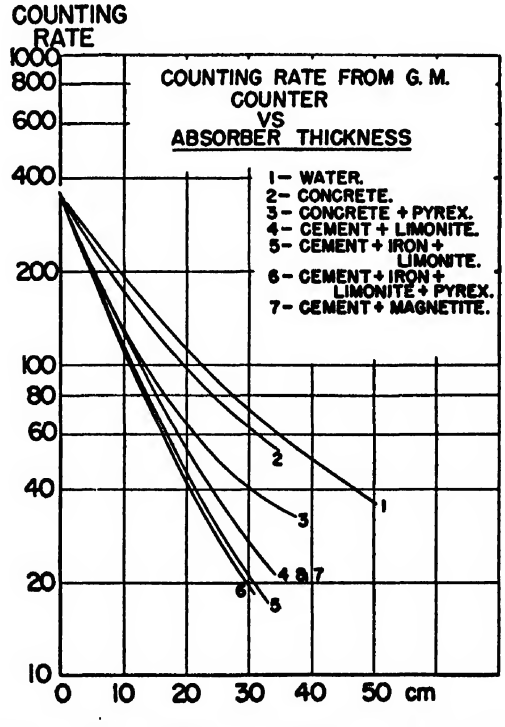


FIG. 11. γ-ray attenuation curves. Abscissa shows the thickness of absorber, ordinate the normalized counting rate of a G-M counter.

TABLE II. Approximate shield thickness for a required attenuation.

		Thickness for neutrons (cm)					Thickness for gamma-rays (cm)				
A[a]	n[b]	Water[c]	Concrete[c]	Magnetite[c]	Limonite[c]	Scrap iron[c] and limonite	Water[c]	Concrete[c]	Magnetite[c]	Limonite[c]	Scrap iron[c] and limonite
10^{-1}	3.33	18.3	35.6	28.3	16.0	17.5	49.4	44.4	30.0	30.0	26.3
10^{-2}	6.66	36.7	71.4	56.6	32.0	35.5	98.8	88.8	60.0	60.0	52.6
10^{-3}	9.99	55.0	107.0	85.0	48.0	53.5	148.2	133.2	90.0	90.0	78.9
10^{-4}	13.32	72.6	142.5	113.2	64.0	71.2	197.6	177.6	120.0	120.0	105.2
10^{-5}	16.65	91.5	178.1	141.5	80.0	89.0	247.0	222.0	150.0	150.0	131.5
10^{-6}	19.98	110.0	214.0	169.8	96.0	107.0	296.4	266.4	180.0	180.0	157.8
10^{-7}	23.31	128.2	250.0	198.1	112.0	125.0	345.8	300.8	210.0	210.0	184.1

[a] A is attenuation factor. [b] n is number of half-intensity values. [c] Columns give thickness in cm to reduce initial intensity by the facto

absorption curve, Rush[18] observed in water with Ra+Be neutrons a half-value thickness of 5 cm is obtained. For plane geometry, Riezler (see Jensen and Ritter[9]) obtained a half-value thickness of 4.9 cm. (For our Be+p neutrons: $D_{th}=5.5$ cm) Jensen and Ritter[9] found a half-value thickness for Ra+Be neutrons slowed down to the In resonance in standard concrete: $D_{In}=10.6$ cm. (For Be+16-Mev p neutrons: $D_{th}=10.7$ cm.) This is in agreement with the high energy part of the spectra obtained for the neutrons from Ra+Be[12] and for the neutrons from the Be^9+p reaction.

It is apparently difficult to increase the density of an absorber without at the same time decreasing the amount of hydrogen available for slowing down neutrons. Since elastic scattering by a heavy nucleus, such as iron, is very inefficient for slowing neutrons it is of interest to investigate the cross section for inelastic scattering and to compare the resulting mean free path for energy loss with that in water. Barshall, *et al.*[19] found about 10^{-24} cm^2 for the inelastic scattering cross section of iron at 3 Mev, and presumably at higher energies the cross section becomes still larger. At 3 Mev the scattering cross section for hydrogen is $2.3\cdot10^{-24}$ cm^2 and decreases as $1/E$ at higher energies. From these cross sections and the composition of a concrete block containing scrap iron and limonite we find that the mean free path for scattering by hydrogen is 18.5 cm, while the comparable number for inelastic scattering by iron is 28 cm. Pure water has a mean free path of about 6 cm for 3-Mev neutrons but for higher energy neutrons the effectiveness of water falls off; so that above about 5 or 6 Mev the iron in our special concrete should be more effective than pure water. To be most effective against high energy neutrons a shield should therefore include heavy nuclei, but in order to slow down neutrons in the range of about a volt up to 3-Mev hydrogen material must be relied upon. These qualitative conclusions are fully borne out by our experiments.

Table I summarizes the data obtained from our experiments. Column 2 gives the amount of water in g/cm^3 calculated (1), from the water added to the mixture, (2) the moisture content of the components and (3), the water of crystallization of the limonite. Column 3 shows the weight of scrap iron+iron in the limonite per cm^3 of material, column 5 the half-value thickness for the neutron attenuation curves obtained with the Ag thermal neutron detectors. The age of thermal neutrons is calculated in column 6. For all concretes, values for the age are appreciably larger than for water. This is probably due to neutrons of about 1 Mev, at which energy the mean free path for a hydrogen collision is still relatively large and inelastic scattering by the iron improbable. This is especially the case in the cement+magnetite test block.

In Table I columns 9 and 10 give the transmission of a 3-ft. wall calculated for two conditions:

1. For neutrons by the expression:

$$T=\frac{\int_{90}^{\infty} I_{th}(r)dr}{\int_{0}^{\infty} I_{th}(r)dr}$$

(For large thickness of absorber r, and exponential attenuation with half-value thickness D, this may be approximated by $T=2^{r/D}$.)

2. For the G-M counter: The absorption curves for the counter are extrapolated with the absorption coefficient obtained from the slope of the curve at $r=30$ cm.

Column 9 shows that for our neutron spectrum the limonite mixtures are as effective as water in stopping neutrons and that compared to ordinary concrete a 3-foot wall of limonite+scrap iron mixture is 280 times more effective against neutrons. Column 10 shows that against γ-rays the limonite plus scrap iron wall is superior to ordinary concrete by a factor of 20, and in comparison with water there is a 60-fold advantage. It is probable that the superiority of iron+limonite would be even more marked at higher neutron energies.

Although boron is frequently suggested for use in shields on account of strong non-radiative capture of neutrons it is apparent from our results that little is to be gained in this direction with respect to either thermal neutron attenuation or suppression of γ-rays.

The concrete+Pyrex absorber has been tested at 90° to the proton beam. A half-value thickness of 5.9 cm

[18] J. H. Rush, Phys. Rev. 73, 271 (1948).
[19] Barshall, Battat, Bright, Graves, Jorgenson, and Hanley, Phys. Rev. 72, 881 (1947).

TABLE III. Measurements on completed shield.

Position in Fig. 12	Shield thickness cm	Ag det. Actual shield	Ag det. Predicted	Counter Actual shield	Counter Predicted
1	0	1[a]	1	1	1
2	30	0.15	0.10	6.8×10^{-3}	5.9×10^{-3}
3	60	$<5\times10^{-4}$	2.4×10^{-3}	3.2×10^{-3}	5.5×10^{-3}
4	90	—	4.9×10^{-4}	5.0×10^{-4}	5.0×10^{-4}

[a] As in the block tests it was necessary to protect the zero absorber silver disk against thermal neutrons coming from the cyclotron by interposing a cadmium filter. The activity of the unshielded detector was 1.77 times larger.

was obtained instead of 7.4 cm in the forward direction, showing that the lower average neutron energy at 90° does indeed lead to larger absorption.

VI. PRACTICAL CONSIDERATIONS

To shield a large installation is always expensive; so practical considerations such as cost, mechanical properties, availability of raw materials and ease with which the shield material can be applied to the actual problem may take precedence over shielding efficacy.

From the mechanical strength standpoint, all the various concrete mixtures showed about the same ultimate compression strength. For example, our tests gave a 28-day strength for ordinary 1-2-3 concrete of 3675 pounds per square inch, while limonite concrete ran 3600 and limonite plus scrap iron about 3000. The early setting properties were normal. Concrete made from scrap iron and limonite is heavy and more difficult to handle in mixing trucks and more difficult to place without voids, but the special problems raised by it are easily solved. The high density of this form of concrete requires sturdy pouring forms and vigorous rodding.

Strength tests showed that the limonite mixtures are adequate for the construction of a radiation shield. An ultimate compression test of cement+iron+limonite cylinders 14 days after casting gave a value of 2193 lb./in.^{2}. For the standard concrete mixture a value of 2210 lb./in.^{2} was observed 7 days after casting.

Limonite ore is widely available, but care must be exercised in specifying high density since some limonite is soft and crumbly. Ore used in our experiments was composed largely of hard, dense lumps ranging from $\frac{1}{8}$ in. to 4 in. to 5 in. A laboratory type crusher reduced the ore to a range of sizes comparable with customary concrete practice. When making up the limonite for incorporation with scrap iron, the ore was crushed somewhat finer than for a pure limonite concrete. A point to notice in specifying limonite is that the amount of absorbed moisture varies considerably with mining conditions and in consequence the amount of water which must be added to make satisfactory concrete will vary somewhat from ordinary concrete practice.

Scrap iron, in the form of small punchings, should be not over $\frac{3}{4}$ in. in size with $\frac{1}{2}$ in. pieces being considerably easier to work. The iron must be grease free and preferably mildly weathered.

Of interest is the comparative cost of the various concretes. Ordinary concrete runs about \$10–\$15 per cubic yard, or about \$10.00 per ton delivered. Limonite ore can be had for \$5.00 per ton at the mine with freight charges increasing the cost to about \$17.00 per ton of ore. Crushing costs are fairly nominal and depend largely on the availability and cooperativeness of local gravel companies. Mixing and delivering of the special concretes, when handled by a commercial concrete firm, may cost as much as the raw materials. Because of the above uncertain factors, it is difficult to quote a reasonable price for limonite concrete, but probably \$40.00 per cubic yard is on the high side. Still more difficult to estimate is the price of scrap iron plus limonite concrete because scrap iron in the most desirable shape is hard to find. The nominal price of scrap iron is presently* \$30–\$40 per ton. Mixing charges are higher for the more dense concretes because of the smaller volume which the mixing trucks can carry. A fair price estimate would be perhaps \$90.00 per cubic yard.

Table II presents the data of Table I in a way which permits cost comparison of different materials for a given attenuation factor. The wall thickness to produce a given attenuation was calculated from the measured half-value thickness on the assumption that there is no change in penetrating power of the radiation for a thickness large compared with 50 cm. Also, for small attenuation the numbers quoted are not accurate because the absorption law is not a simple exponential, e.g., there are *more* thermal neutrons behind 10 cm of concrete than with no concrete at all. With these reservations in mind, one may conclude that to reduce the thermal neutron flux by a factor of one million either water or limonite concrete of one-meter thickness

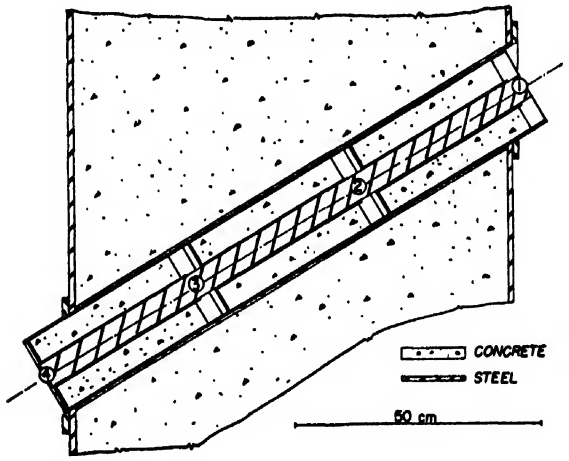


FIG. 12. Cross section through actual cement+limonite+scrap iron shield, showing the port hole, which is directed towards the Be target. The numbers 1 to 4 refer to the positions where the silver detectors and G-M counters were placed to test the actual performance of the shield.

* 1948.

will suffice. Ordinary concrete would have to be about twice as thick. On a cost basis, water is by far the cheapest with straight concrete next and the special concretes much more expensive. On the other hand, a meter of water gives only 10^{-2} attenuation for radiation detectable by a G-M counter while a meter of limonite concrete gives a factor of 10^{-3}, or better, and the scrap iron concrete gives a factor of 10^{-4}. In many installations it is very inconvenient to have a 6-foot wall of water or concrete around the equipment, and in such cases the item of cost is overbalanced by the need for compactness.

Where space and convenience are not at a premium, ordinary concrete is probably the best shielding material. In those regions where limonite, or even magnetite, are readily available and where the transportation, crushing and mixing costs can be kept down it is worth while to use these more costly concretes. On the other hand, the scrap iron concrete is so costly that its use is indicated only where the greater compactness is worth the added price.

VII. PERFORMANCE OF THE COMPLETED SHIELD

Measurements on the finished cement, limonite, and scrap iron shield which has a thickness of 3 feet proved that the neutron and γ-ray attenuation agreed with the extrapolation of the results obtained with the absorber blocks. Figure 12 shows a cross section through that part of the wall in which a port hole has been inserted. The channel is parallel to the neutron flux and in the forward direction of the proton beam incident on the Be target. Three concrete slabs fill the channel. For experimental purposes a 2-in. hole is left in the slabs, which can be closed by a steel rod. For the attenuation measurements, silver disks or a Geiger counter can be placed between the slabs in the usual way.

In Table III the results obtained with the shield are compared with the attenuation expected from the previous experiments. The activity of the Ag detector at position 3 of Fig. 12 was just above the background with the cyclotron off. Only a lower limit for fthe attenuation about 5×10^{-3} could be established at that position; this compares with the value of 2.4×10^{-3} expected from the absorption experiments.

The absorption of the γ-rays appears to have been a little higher than expected, 3.2×10^{-3} against 5.5×10^{-3} at 60 cm. Going from 60 to 90 cm at the outside of the shield the counting rate is only decreased by a factor $\frac{1}{3}$. This is caused by γ-rays transmitted through the doors in the shield which at the time of these measurements are filled with water.

The counting rate for a Victoreen Mica window type counter placed at position 4 of Fig. 12 is 150 counts a minute when a target of Be is bombarded with about 1 μa of 16-Mev protons.

The agreement between observed and empirically predicted attenuation shows that results obtained by our small block tests are sufficiently reliable to permit the design of large shields.

We are indebted to several members of the Princeton faculty for advice on concrete and iron ore problems. In particular, we wish to thank Professor H. F. Winterkorn for his help on concrete problems and Professors A. F. Buddington and E. Sampson for advice on limonite ores. The details of concrete mixing and structural testing were carried out by Professor J. G. Hendrickson of the Engineering School and Mr. B. Setzer of the cyclotron group.

A Bending Method for Measurement of the Tensile Strength and Young's Modulus of Glass Fibers

DAVID SINCLAIR
Johns-Manville Research Center, Manville, New Jersey
(Received August 26, 1949)

The tensile strength and Young's modulus of Johns-Manville experimental drawn glass fibers were measured by a bending method which consists of twisting a loop in a fiber and pulling the ends until the loop breaks. The tensile strength observed by this method was two to three times higher than observed by the conventional straight fiber method, while Young's modulus was practically unchanged. These results may be accounted for by the presence of flaws, since the probability is small that the weakest flaw will occur in the loop, while the straight fiber will always break at the weakest flaw. The apparatus used, a modified Chainomatic balance, is described and the equation of the looped fiber is derived. The method is used to measure the maximum tensile stress and strain attained at the point of minimum radius of curvature of the loop as well as the breaking tensile stress and strain at the actual break point of the fiber. The stress at the break point was found to be as small as one-half the maximum attained in the loop, showing that the strength of the fiber was not uniform.

INTRODUCTION

THE tensile strength of glass fibers is known to vary from a few thousand to two or three million lb./sq. in.[1] depending upon the diameter and length of the fiber, the method of manufacture and amount of handling before test, the age of the fiber and the elapsed time during the test. This wide variation in tensile strength has been supposed to be caused by flaws[1,2] which weaken the fiber at numerous points. The idea is supported by the fact that the tensile strength of fibers increases with decreasing length,[3] since the number of flaws decreases with decreasing length. A very short fiber might therefore be free of any weak flaws and consequently show an unusually high tensile strength.

This paper describes a bending method which provides a measurement of both the breaking stress and strain over a very short length of fiber. The tests were run on Johns-Manville continuous glass fibers made from a plate glass type composition. The method consists of twisting a small loop in a fiber 1 or 2 in. long and pulling the ends until the loop breaks. The tension required to break the loop and the size of the loop at the instant of break are used to calculate the tensile strength at the break point of the loop and the tensile stress attained at the point of minimum radius of curvature of the loop. In addition the extension at these points may be calculated, yielding Young's modulus.

The looped fiber is equivalent to a very short straight fiber since the looped fiber is strained to its breaking point over a length of 1 mm or less. The method is particularly suited to materials such as glass which break under small deformation. In the straight fiber method the extension is too small to be observed conveniently, and short fibers tend to break at the supports. In the looped fiber method the large reduction in the size of the loop during a run magnifies the strain so that it can be readily measured.

Murgatroyd[4] used a somewhat similar method to obtain the breaking strain but not the stress of bent fibers. The fiber was bent in the form of a hairpin between the jaws of a miniature vise which was closed until fracture occurred. A maximum extension of about three percent was observed in 25-micron diameter fibers.

METHOD

The apparatus consists of a Chainomatic balance to which is attached a recording pen, and a 22 speed motor-driven Mono-Drum and loading device. Figure 1 shows the recording pen, P, attached to the chain weight drive, W, which is operated by hand in the usual manner. The drum, D, 10 in. long and 6.25 in. in diameter (50 cm circumference), is motor driven at a constant speed of 1 r.p.m., always in the same direction (from right to left under the pen). The looped fiber is hung from the left pan hook of the balance, and its lower end drawn vertically downward at a constant speed of 0.5 in. per min. by the loading rod, R, which is geared to the drum. A gear shift, G, is used to reverse the direct of motion of the loading rod in preparation for a new test. The magnetic damper, M, provides aperiodic damping of the balance.

In order to mount a fiber for test, one end of a straight fiber 1 or 2 in. long is cemented to a small wire hook which is then hung on the left pan hook. A suitable cement is Picein or low melting point De Khotinsky which is melted by means of an electrically heated wire placed against the wire hook. The lower end of the fiber is then cemented to a piece of straight wire held in a pin vise, V, mounted in the loading rod. The pin vise is free to be rotated or raised in the loading rod in order to facilitate formation of the loop. The cemented ends of the fiber are then straightened by applying a slight tension of about 10 mg with the chain weight, and heating each cemented end enough to soften the cement.

With the slight tension still applied to the straight

[1] C. J. Phillips, *Glass the Miracle Maker* (Pitman Publishing Corporation, New York, 1948), Chapter 4.
[2] F. O. Anderegg, Ind. Eng. Chem. 31, 290 (1939).
[3] J. J. Bikerman and G. H. Passmore, *The Glass Industry* 29, 144 (March, 1948).
[4] J. B. Murgatroyd, J. Soc. Glass Tech. XXVIII, 368, 388 (December, 1944).

fiber the loading rod is driven up or down (and simultaneously the drum is rotated) until the balance needle is at the zero mark. This point is then located on the drum chart (Fig. 2) by a short vertical line near the top of the chart, drawn by returning the chain weight (with attached pen) to zero. This vertical line locates on the chart the abscissa corresponding to the position of the lower end of the straight fiber. The loading rod is then driven up 0.5 in. corresponding to one revolution of the drum and an abscissa of 50 cm on the chart. A loop is now formed in the fiber by turning the pin vise $1\frac{1}{2}$ to 2 revolutions, raising it until the loop snaps into place and then lowering it to its original position (determined by a shoulder on the pin vise).

The vertical line on the chart now corresponds to the position of the lower end of the looped fiber, which is 0.5 in. above the original position of the lower end of the straight fiber. The equations show (see below) that the elevation, L, of the lower end of the looped fiber above the lower end of the straight fiber is proportional to the minimum radius of curvature, R_m, of the loop, i.e.,

$$L=8R_m. \tag{1}$$

The horizontal line on the chart, drawn during the one revolution of the drum, gives the position of zero force on the fiber. The initially large loop exerts a small downward force, T_0, on the left arm of the balance, which should be compensated by lowering the chain weight until the balance needle is again at the zero mark. The looped fiber is now ready for test.

The test is run by driving the loading rod downward while continuously adding chain weight to keep the balance needle at zero. When the loop breaks, no further weight is added, and the drum is allowed to run slightly to record the breaking tension. The wire carrying the lower portion of the fiber is then removed from the pin vise and saved for measurement of the fiber diameter.*

The tension, T, of the fiber at any point on the curve is proportional to the vertical distance from the point to the horizontal zero line. The constant of proportionality is determined by calibration by weighing. Different ranges were obtained by hanging the chain from different points on the balance arm. The minimum radius of curvature, R_m, is proportional to the horizontal distance, L, from the point to the vertical starting line, measured in the direction of travel of the pen. As described above, one horizontal cm on the chart corresponds to 0.01-in. vertical displacement of the lower end of the fiber.

It is shown below that the maximum tensile stress, σ, in the fiber is:

$$\sigma=\frac{16TR_m}{\pi r^3}=\frac{16TL}{\pi d^3}, \tag{2}$$

and the corresponding extension, e, is:

$$e=r/R_m=4d/L, \tag{3}$$

where r is the radius and d the diameter of the fiber.

Young's modulus of elasticity, E, is:

$$E=\sigma/e=(4TL^2/\pi d^4). \tag{4}$$

The ultimate tensile stress and strain are obtained by measurement from the break point on the chart record.

Some practice is required to keep the balance in equilibrium during a test. The rate of increase of tension is at first very low and then increases until it becomes quite rapid as the break point is approached. The operation is facilitated by observing the balance needle through a magnifying lens (not shown in Fig. 1) of the type manufactured for accurate weighing. The curve obtained is slightly wavy, but the fluctuations are small (usually ± 1 mg), particularly as the loop approaches the break point. Figure 2 shows a representative curve redrawn to reduced scale.

The tension required to break the loop in a 20-micron diameter fiber is of the order of 50 mg. Observations of the looped fiber test performed on a completely automatic, electronically operated Chainomatic balance now available on the market, showed severe hunting of ± 10 mg, resulting in a saw-toothed curve on which the location of a point, in particular the break point, was uncertain. A completely automatic instrument of improved stability has recently been developed in this laboratory and is now being tested.

Equations (2) and (3) give the maximum value of tensile stress and strain, respectively, attained in the loop at the point of minimum radius of curvature, which occurs at the base, O, of the loop (see Fig. 4). Measurements of the break point, as described in the following paragraphs, have shown that the fiber frequently breaks at some other point in the loop where the stress and strain are less than at the base, showing that the

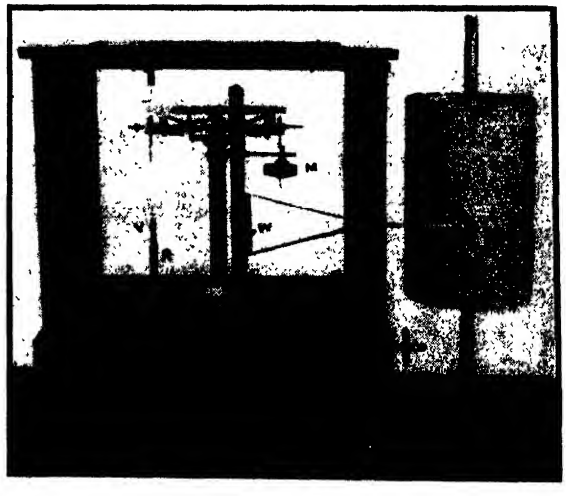


FIG. 1. Semi-automatic fiber balance.

* The diameter was measured with a microscope by placing the broken end of the fiber on a slide in a drop of oil of refractive index 0.05 greater or less than that of the fiber.

strength of the fiber was not uniform. This non-uniformity is not due to variations in the diameter of the fiber. Measurements of the diameter of 20-micron diameter glass fiber samples over a length of several millimeters from the broken end have shown the diameter of any one sample to be constant to two percent. Such a variation in the diameter would cause a variation in the calculated tensile strength of six percent, but the loop has been observed to break at a point where the tensile stress was 50 percent below the maximum attained at the base of the loop.

The measurements of the loop break point were made entirely manually with a preliminary form of apparatus, shown in Fig. 3. The fiber was mounted horizontally with one end fixed in position and tension applied at the other end by a light, circular spring attached to a graduated mechanical slide. The diameter, D, and the position of the loop were observed at the instant of break by means of a low power microscope and micrometer eyepiece. A reading of the mechanical slide was taken corresponding to the straightened fiber with zero tension in the spring, and another at the loop break point. The difference, A, between these two readings is equal to the displacement of the end of the spring attached to the mechanical slide, as shown in Fig. 3.

The displacement, L, is equal to the oppositely directed displacement of the other end of the spring to which the fiber is attached. The equations show that the displacement, L, is proportional to the diameter, D, of the loop (see Fig. 4), i.e.,

$$L=3.75D. \tag{5}$$

Consequently, the total extension of the spring and therefore the tension on the fiber at the instant of break is proportional to $L+A=3.75D+A$. The constant of proportionality was determined by calibration with a Chainomatic balance.

The equations also show that if the fiber breaks at the base of the loop, the remaining portion of the straight fiber (attached to the fixed end) will extend a distance $s=\frac{1}{2}L$ beyond the center of the loop. If the fiber breaks at a point other than the base, the straight fiber will extend a distance $s\pm l$ beyond the center of the loop. Here l is the distance along the fiber from the base of the loop to the break point. Equation (23) shows the tensile stress in the loop as a function of the break point.

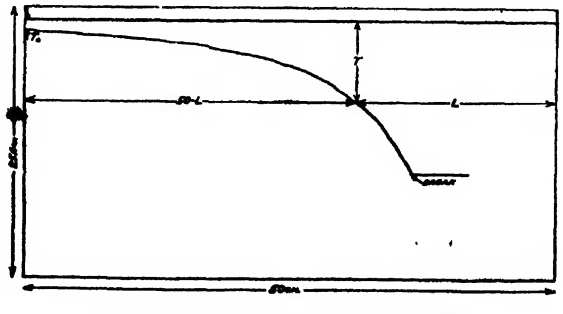


FIG. 2. Chart record of looped fiber test.

After the loop was broken, the fixed portion of the fiber was straightened by wetting the portion attached to the spring and then touching the two portions together along a considerable part of their length. The two portions then clung together so that a slight tension sufficient to straighten them could be applied by the spring. The location of the end of the fixed portion, relative to the recorded position of the loop at the time of break, was read on the micrometer eyepiece. This measurement yields the value of l.

From the above measurements of D and A and the fiber diameter, the tensile stress at the base of the loop can be calculated from Eq. (2). The tensile strength at the actual break point is obtained from Fig. 5, a plot of Eq. (24).

RESULTS

Measurements were made on continuous glass fibers when dry (i.e., not lubricated) and when lubricated with oil applied during the drawing process. Table I gives a summary of the results obtained when using the semi-automatic balance on looped fibers, and a Scott Serigraph on straight fibers. Column 1 shows the type of fiber, column 2 the number of samples tested, column 3 the mean fiber diameter, and column 4 the average deviation from the mean (a.d.). Column 5 shows the arithmetic mean tensile strength, σ_m (at the minimum radius of curvature), column 6 the highest, and column 7 the lowest value of tensile strength and column 8 the average deviation from the mean. Columns 9 to 12 show the corresponding values for the percent extension (also calculated at the minimum radius of curvature), columns 13 to 16 the corresponding values for Young's modulus and column 17 the method.

It is seen that the tensile strength of the dry looped fibers of 20 microns diameter is in average more than twice that of the straight dry fibers. For the lubricated fibers of about 12 microns diameter, the tensile strength of the looped fibers is more than 2½ times that of the straight fibers. No comparison of the extension and Young's modulus could be made between the looped fibers tested on the balance and the straight fibers tested on the Scott Serigraph since the latter apparatus is not sufficiently sensitive to record the small deformation occurring in glass. Previous measurements on straight fibers,† showed the ultimate tensile strain in straight fibers to be one-half to one-third that attained in looped fibers. This ratio is roughly equal to the ratio of tensile strengths obtained by the two methods, so that Young's modulus is about the same for looped and straight fibers.

The straight fibers show an increase in tensile strength with decreasing fiber length. The dry fibers below 1 in. in length are in average 18 percent stronger than those over 1 in. long, and the shorter lubricated fibers are 34

† The straight fiber was broken by applying tension with a calibrated coil spring, and the extension observed with a microscope.

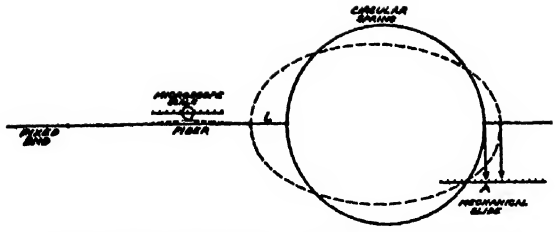


FIG. 3. Schematic diagram of loop break point test.

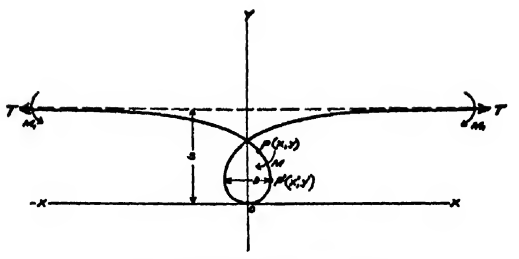


FIG. 4. Diagram of looped fiber.

percent stronger. Straight fibers about 1 mm long would presumably have about the same tensile strength as the looped fibers, provided such short straight fibers could be mounted free of injury or initial strain. This was found to be impossible in practice, the losses of fibers even $\frac{1}{4}$ in. long being excessive. The fiber frequently breaks at one of the cemented ends, or the shock at break causes both portions to fly off since the breaking force of a straight fiber is about 1000 times that of a looped fiber. The losses in the straight fiber test were high even for fibers 1 or 2 in. long. The 39 successful tests of straight fibers over 1 in. long were obtained from a total of 65 tests, while only 38 successful tests of fibers less than 1 in. long could be salvaged from a total of 110. The losses in the looped fiber method were negligible.

No trend in tensile strength with fiber diameter could be observed, because the variation due to flaws was so large. When the tensile strength of the looped dry fibers was plotted against fiber diameter (which ranged from 16 to 32 microns) a typical shot-gun pattern was obtained. The same result was obtained with the lubricated fibers whose diameter ranged from 9 to 16 microns.

Observations of the break point of the loop were made on nine fibers measured with the manually operated spring (Fig. 3). The results, expressed as the ratio of the tensile strength at the break point to the tensile stress at the base, were as follows: 0.999, 0.970, 0.999, 0.890, 0.999, 0.515, 0.997, 0.540, 0.988. These results show that six of the fibers broke very near the base (see Fig. 5), while two of them broke where the tensile strength was little more than one-half the maximum tensile stress attained at the base. In both of the latter tests the diameter of the loop was 0.66 mm, the minimum radius of curvature 0.31 mm and the break occurred within the loop at a distance of about 0.7 mm from its base where the radius of curvature was twice the minimum, or 0.62 mm.

These results show that a region of glass fiber about 1 mm in length can have a tensile strength at least twice as great as that of a nearby region. Additional measurements would probably show a greater variation in the strength of neighboring regions of a fiber. They were not made since this type of measurement is very tedious, and they appear to be unnecessary. The results show nearly a fourfold difference in tensile strength of looped, dry fiber samples. It is to be expected that a sufficiently large number of measurements of the loop break point would eventually show an equally large difference in neighboring regions of some one sample.

DISCUSSION

The results show that the tensile strength of glass fibers increases two to three times when their length is reduced about 75 times. Under the conditions of experiment the variation in the other factors mentioned in the introduction was sufficiently small to have little effect. Although the fibers had a two to one range in diameter the mean diameter was so nearly the same in both types of test that the effect of varying diameter was averaged out. The fibers were taken at random from the same batch and all tests were run at a room temperature of about 75°F and a relative humidity of 45 percent.

Since the fibers were two or three years old and had been roughly handled the observed tensile strength is not representative of the type of glass tested. However, the relative values of tensile strength obtained in the looped and straight fiber tests are in agreement with Anderegg's[2] measurements of the variation of tensile strength with fiber length. Anderegg found that a decrease in length of about 32 times resulted in doubling the tensile strength of 13-micron diameter soda-lime glass fibers. The looped dry fibers (equivalent to a straight fiber about 1 mm long) were $2\frac{1}{4}$ times stronger and the looped lubricated fibers three times stronger than the straight fibers of about 3 in. average length.

The reliability of the 74 tests of looped dry fibers was calculated according to ASTM[5] methods, based on a statistical probability of 0.99. The calculations give tensile strength $\sigma = 2.50 \pm 0.12$ lb./sq. in. (standard

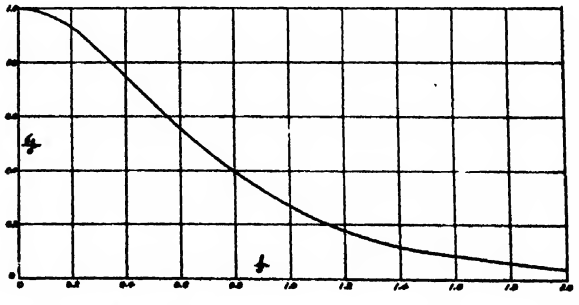


FIG. 5. Tensile stress in looped fiber vs. break point. Plot of Eq. (24).

[5] ASTM Manual on Presentation of Data, Supplement A (April, 1945).

TABLE I. Tests of glass fibers.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Fiber type	No. samples tested	Mean fiber diam. microns	a.d.	Tensile strength 10^4 lb./sq. in. Mean	High	Low	a.d.	Percent extension Mean	High	Low	a.d.	Young's modulus 10^7 lb./sq. in. Mean	High	Low	a.d.	Method
								Looped fibers								
Dry	74	20.4	2.4	2.50	3.68	1.55	0.29	3.06	4.76	1.89	0.43	0.83	1.19	0.47	0.10	Balance
Lub.	31	11.3	1.3	3.23	5.29	2.12	0.58	3.40	5.60	1.86	0.68	0.95	2.05	0.50	0.28	Balance
								Straight fibers								
Dry	23	20.3	3.5	1.11	2.01	0.61	0.29				Over 1 in. long					
	12	19.5	2.0	1.31	1.64	0.92	0.23				Less than 1 in. long					Scott Serigraph
Lub.	16	12.2	1.8	1.01	1.48	0.70	0.21				Over 1 in. long					
	26	13.2	1.7	1.35	2.95	0.56	0.47				Less than 1 in. long					

deviation=0.379) and Young's modulus $E=0.825 \pm 0.042$ lb./sq. in. (standard deviation=0.138).

The difference between loading rates in the two tests was relatively small. The loading rate in the straight fiber test was constant at 0.12 lb./min. The rate of stressing, however, varied between 10^5 and 10^6 lb./sq. in./min. depending on the fiber diameter and Young's modulus. The maximum rate of stressing (i.e., in the outermost layer of the fiber) at any time during the looped fiber test can be obtained by differentiating the tensile stress, σ, with respect to the time: [see Eqs. (3) and (4)]

$$\frac{d\sigma}{dt}=E\frac{de}{dt}=-E\frac{e}{L}\frac{dL}{dt}=-\frac{\sigma}{L}\frac{dL}{dt}. \tag{7}$$

dL/dt is negative and constant at 50 chart cm per min. Therefore:

$$d\sigma/dt=50\sigma/L \text{ lb./sq. in./min.} \tag{8}$$

where L is to be measured in cm on the chart as described above. Thus the rate of stressing increases as the test progresses directly with the tensile stress and inversely with the loop diameter.

For example, in Fig. 2 the ultimate tensile stress in this test was 2.70×10^5 lb./sq. in. and the corresponding value of L is seen to be 13.6, from which $d\sigma/dt=10^6$ lb./sq. in./min. At an earlier time during the test when $L=28.7$, $\sigma=1.3\times10^5$ and $d\sigma/dt=2.3\times10^5$. The rate of stressing is of course very small initially but during the latter half of the loop test the order of magnitude of the rate is the same as in the straight fiber test. Anderegg[2] found that a fourfold increase in the rate of stressing of 44-micron diameter soda-lime glass fibers produced no significant increase in the tensile strength. Baker and Preston[6] found that a tenfold increase in the duration of a constant load caused a 15 percent decrease in the bending strength of ¼-in. diameter soda-lime glass rods.

As already described, observation of the actual break point of the loop showed non-uniformity of strength over a very short section of a fiber. In most tests the tensile strength was calculated at the minimum radius of curvature, although the loop may have broken at some other point where the ultimate stress was less. In such cases the calculated value of stress was attained at the minimum radius of curvature without breaking at that point, so that this value represents a minimum stress for this point of the fiber. Also, it may be mentioned that the straight fiber losses occurred equally at both high and low loads so that the values reported are a fair average of the tensile strength obtainable in straight fiber tests.

DERIVATION OF EQUATION OF LOOP

Figure 4 shows a diagram of the looped fiber drawn with the base of the loop, the point of minimum radius of curvature, at the origin, O. The loop is maintained by the equal and opposite bending moments, M_1, and tensions, T, exerted at each cemented end by the supports. T acts along a line parallel to the X axis and at a distance, s, from it. In addition there is an equal and opposite torque at each end, exerted about the line through T, to keep the loop from untwisting. As shown below this torque is so small it may be neglected. The looped fiber is symmetrical about the Y axis.

At any point, $P(x, y)$, of the fiber, let the bending moment be M. The portion of the fiber between P and the left cemented end is maintained in equilibrium by the clockwise moment, M, exerted by the remainder of the fiber, and the counterclockwise moments, M_1, and $T(s-y)$.

Therefore,

$$M=M_1+T(s-y). \tag{9}$$

The bending moment $M=EI/R$, where R is the radius of curvature of the fiber at P, E is Young's modulus of elasticity and I is the moment of inertia of the cross section of the fiber.

When the fiber is long compared to the diameter of the loop the bending moment, M_1, at the end of the fiber is negligibly small (see below). Therefore, writing R in

[6] T. C. Baker and F. W. Preston, J. App. Phys. 17, 170 (1946).

terms of the differential coefficients, Eq. (9) becomes:

$$\frac{1}{R}=\pm\frac{pdp/dy}{(1+p^2)^{\frac{1}{2}}}=b(s-y), \tag{10}$$

where $p=dy/dx$ and

$$b=T/EI. \tag{11}$$

The positive sign on the left side of Eq. (10) applies from 0 to P' (at which point $dy/dx=\infty$), where R is positive; and the negative sign applies from P' to the end of the fiber, where R is negative.

Integrating Eq. (10) gives:

$$\pm 1/(1+p^2)^{\frac{1}{2}}=\tfrac{1}{2}by^2-bsy+c. \tag{12}$$

At the origin $y=0$ and $dy/dx=p=0$. Therefore, using the positive sign, $c=1$.

At the cemented end $y=s$ and $dy/dx=0$. Therefore, using the negative sign:

$$b=4/s^2 \tag{13}$$

so that Eq. (12) becomes:

$$\pm 1/(1+p^2)^{\frac{1}{2}}=(2y^2/s^2)-(4y/s)+1. \tag{14}$$

Integrating Eq. (14) gives:

$$x=\pm s\left(\frac{2y}{s}-\frac{y^2}{s^2}\right)^{\frac{1}{2}}\mp\frac{s}{4}\log\frac{1+(2y/s-y^2/s^2)^{\frac{1}{2}}}{1-(2y/s-y^2/s^2)^{\frac{1}{2}}}. \tag{15}$$

The upper signs apply to the left half of the curve and the lower signs to the right half. Equation (15) has been plotted and its form found to fit closely to the observed form of the loop as shown below.

The measurement of the tensile strength by the method described, requires the length of the fiber, l, which is found to be:

$$l=-\frac{s}{4}\log\frac{1+(2y/s-y^2/s^2)^{\frac{1}{2}}}{1-(2y/s-y^2/s^2)^{\frac{1}{2}}}. \tag{16}$$

Equation (16) gives the length of the fiber, measured from the origin along the fiber to a point having the ordinate y.

Equation (2) gives the tensile strength in terms of L, the total length of the fiber minus the distance between the ends of the looped fiber. The length of the fiber between the points (x, y) and $(-x, y)$ is $2l$, given by Eq. (16). The linear distance between these two points is $2x$, given by Eq. (15). Subtracting Eq. (15) from Eq. (16) and multiplying by 2 gives:

$$2l-2x=2s[(2y/s)-(y^2/s^2)]^{\frac{1}{2}} \tag{17}$$

for negative values of x on the left half of the curve and for positive values of x on the right half of the curve.

The difference, $2l-2x$, is equal to the distance L when $y=s$, the ordinate of the end of the fiber. Although, from Eqs. (15) and (16) the abscissa of the end of the fiber and the length of the fiber are both infinite, their difference is finite as shown by Eq. (17). In any practical case the equations apply when the length of the fiber is more than 10 times the diameter of the loop.

From Eq. (17) when $y=s$,

$$2l-2x=2s=L. \tag{18}$$

Equation (4) for Young's modulus of elasticity can now be obtained from Eqs. (11), (13), and (18), noting that for a fiber of circular cross section of diameter d, the moment of inertia $I=\pi d^4/64$.

In the outermost layer of the fiber at a point where the radius of curvature of the loop is R, the extension $e=r/R$.

The minimum radius of curvature, R_m, which occurs at the base of the loop, is obtained from Eq. (10) when $y=0$, i.e.,

$$R_m=1/bs=s/4=L/8 \tag{19}$$

so that the maximum tensile strain in the fiber is:

$$e=r/R_m=8r/L=4d/L \tag{20}$$

corresponding to Eq. (3).

Equation (2) for the maximum tensile stress attained in the loop is given by the product of Eqs. (3) and (4).

The break point of the loop is obtained with the aid of Eq. (18) which shows that the length of the looped fiber is $2s$ greater than the distance between its ends. This means that if the looped fiber were broken at the base (the origin), each straightened portion would extend a distance, s, beyond the Y axis.

Suppose the fiber broke at a distance, l, from the origin, where l is given by Eq. (16). Solving Eq. (16) for y gives:

$$y=s\pm 2se^{(2l)/s}/(e^{(4l)/s}+1). \tag{21}$$

From Eq. (10) when $y\neq 0$, the strain at any point other than the base is:

$$e=(2d/s^2)(s-y). \tag{22}$$

Therefore from Eq. (4), (21), and (22) the maximum tensile stress, σ_l, at a point in the loop a distance, l, from the origin is:

$$\sigma_l=\frac{32TL}{\pi d^3}\,\frac{e^{(2l)/s}}{e^{(4l)/s}+1}. \tag{23}$$

From Eqs. (2) and (23) the ratio of the tensile strength at the break point to the tensile stress at the base is:

$$\sigma_l/\sigma=\frac{2e^{(2l)/s}}{e^{(4l)/s}+1}. \tag{24}$$

Figure 5 shows a plot of Eq. (24) for values of l/s from 0 to 2.

In making the measurements of the loop break point the diameter, D, of the loop is measured across the loop between the points, $y=y'$, where $dy/dx=\infty$ (see Fig. 4). Therefore from Eq. (14):

$$y'=s(1-1/\sqrt{2}) \tag{25}$$

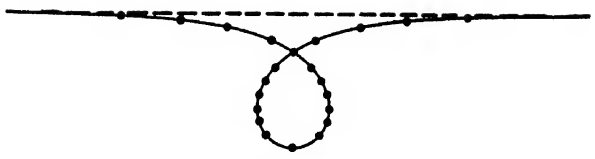


FIG. 6. Photograph of glass fiber loop. Full line—photograph. Circles—calculated points.

so that from Eq. (15), $x'=0.2664s$. Therefore, since $x'=\frac{1}{2}D$

$$D=0.5328s. \tag{26}$$

The assumption that M_1, the moment at the cemented ends, is zero is found to be justified by observation. Since $M=EI/R$, this assumption means that the fiber is straight at the ends. Also from Eq. (15) when $y=s$ this assumption requires a fiber of infinite length. Practically this means that the fiber should be long compared to the size of the loop.

These conditions have been found to be closely maintained in the loops observed. The end portions of the fiber lie along the same straight line to a considerable distance from the cemented ends. (See Fig. 6.) With a fiber 2 cm long the cemented end is located at a distance from the loop of at least 10 times its diameter.

At any point in the fiber, the shearing force, $F=dM/dy=-T$ [from Eq. (9)]. The maximum shear stress occurs at the point P' (Fig. 4) where the shearing force acts perpendicular to the axis of the fiber. At this point the average shear stress, $t=T/\pi r^2$. For a fiber of 10 microns radius and a breaking tension of 50 mg the average shear stress is 225 lb./sq. in. The maximum shear stress, which occurs at the center[7] of the fiber is 310 lb./sq. in. This is negligible compared to the tensile strength of 10^5 lb./sq. in. or more.

The maximum shear in the fiber resulting from the one and one-half to two turns given the fiber to form the loop is produced by one-half to one turn, since one turn is taken out when the loop forms. For a 10-micron radius fiber about 2 cm long the maximum shear strain produced by one turn is 0.3 percent. In most tests one-half residual turn is sufficient to maintain a loop in a fiber 4 or 5 cm long, for which the torsional shear would be less than 0.1 percent.

The one-half turn remaining in a fiber is maintained by equal and opposite couples, C, acting about the cemented ends. The couple, $C=fks$, where f is the normal force between the two halves of the loop at the cross-over point, and ks is the ordinate of the cross-over point. k is a constant $=0.708$. For constant torque in the fiber, C is constant so that the normal force $f=$ const. $\times 1/s$. It was observed that the resulting friction force at the cross-over point was appreciable at the beginning of a test. However, toward the end of the test the friction becomes negligible compared to the tension since the latter increases as $1/s^2$ [Eq. (4)—assuming constant Young's modulus] while the friction force increases as $1/s$. The friction was measured by contracting the loop to near the break and then reversing the direction of the loading rod so as to run the test backward and expand the loop. The friction is added to the tension while the loop is being contracted and subtracted while the loop is being expanded. As a result the return curve tends to lie slightly above the initial curve (of the type shown in Fig. 2), the divergence being significant only at the smaller values of T. In this way the force of friction was found to be about one percent of the tension at or near the break.

[7] S. Timoshenko, *Theory of Elasticity* (McGraw-Hill Book Company, Inc., New York, 1934), p. 290.

EXPERIMENTAL CHECK OF LOOP EQUATION

The ratio, $s/D=1.877$ [Eq. (26)] was measured using glass and Pyrex fibers. The size of the loop was varied by varying the tension over a range of values less than the breaking tension. Using two 20-micron diameter glass fibers, the average of ten measurements gave $s/D=1.887\pm0.005$, 0.5 percent higher than the correct value. Using two 6-micron diameter Pyrex fibers the average of seven measurements gave $s/D=1.870\pm0.015$, 0.4 percent less than the correct value.

A photograph of a 20-micron diameter glass fiber having a loop of 2 mm diameter is shown in Fig. 6. At the edges of the photograph the ends of the fiber are seen to be practically straight and lie on the same straight line (the dashed line drawn to pass through the cemented ends which lie outside the photograph at a distance of one and one-half times the total breadth of the photograph). The ratio $s/D=1.88$ as nearly as can be measured on the photograph. Using the measured value of s, the form of the curve was calculated from Eq. (15). The plotted points lie exactly on the curve as nearly as could be drawn.

ACKNOWLEDGMENT

The writer is greatly indebted to Mr. Paul Greebler of the Physics Section for his assistance in developing the method, and to Mr. H. R. Pittenger of the Physics Section and to Messrs. D. T. Colton, Jr. and Stewart Hayes (formerly of the Physics Section) for running the tests.

Magnetic Field Calculations for Large Cross-Section Cloud-Chamber Coils*

H. W. Koch†
Physics Department, University of Illinois, Urbana, Illinois
(Received September 6, 1949)

A method is described and equations are presented which permit the calculation of the magnetic field characteristics of large cross-section air core coils. The method is applied to the determination of the field non-uniformities on the median plane to be expected with annular cloud-chamber coils of rectangular cross section. The field non-uniformity results are provided in graph form to facilitate the design of cloud-chamber coils.

INTRODUCTION

LARGE, air-core magnetic field coils of the rectangular cross-section type shown in Fig. 1(b) facilitate the study of high energy nuclear particles in the 10 and 100 million electron volt energy ranges. The coils have no windings on the median plane and are used in conjunction with cloud chambers where light and particle beams are admitted on the median plane.

This paper will describe calculations of the magnetic field distributions produced by cloud-chamber coils of large cross-sectional area. Greatest emphasis is placed on the field distribution on the median plane and within the inner radius of the coil. For this region the calculations are presented in a table form. Also, curves are given to allow the determination of the field non-uniformity for almost any rectangular coil configuration with speed and a fair degree of accuracy.

In addition, general equations are derived which permit the evaluation of the absolute magnitude of fields anywhere on the median plane, outside as well as inside the coil bounds. The result is applicable to a wider range of instrument than the cloud chamber: for example, to particle accelerators or to particle analyzers.

To date, the design work on air core coils has been concerned largely with the production of relatively weak, uniform fields for cloud-chamber studies. Since the power requirements were usually small, the field region investigated was chosen close to the axis of the coils. The theoretical methods[1] were based on the assumption that the coil cross section was small compared to the coil radius.

Two general procedures for the small cross-section calculations have been used. In the first method, as employed by Maxwell,[2] the necessary correction due to the cross section was made by the use of Taylor's theorem. The field was expanded to give

$$H=H_0+\frac{1}{24}\left(\xi^2\frac{\partial^2 H_0}{\partial x^2}+\eta^2\frac{\partial^2 H_0}{\partial y^2}\right)+\cdots,$$

where H_0 is the value of the field at the point (x, y) due to a number of coincident circular turns of wire and H is the field at the same point due to the same number of turns of wire spread over a rectangular cross section of axial width ξ and radial depth η.

Gray,[3] Dwight,[4] and Ruark and Peters[5] employ the Maxwell method for calculating fields produced by coils of small cross section. Ference, Shaw, and Stephensen[6] have used Maxwell's method as modified by Lyle.[7]

The second method for handling small cross sections was to integrate the magnetic scalar potential over the area of the coil. Dwight[4] described this method in detail for rectangular cross-section coils. Von Zeipel[8] performed the integration over a cross section which is limited by coaxial cylinders and concentric cones coaxial

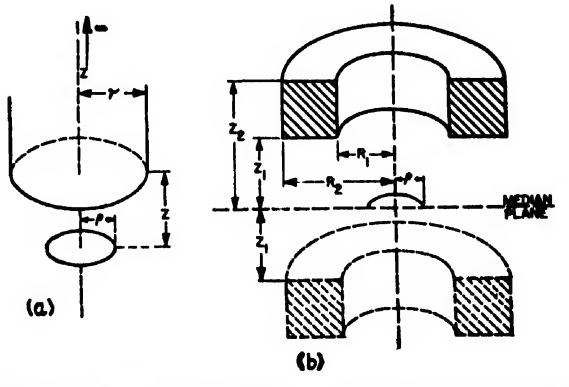


FIG. 1. (a) A semi-infinite solid cylindrical coil of zero core area. (b) The solid rectangular coil above the median plane represents one half of a pair of cloud-chamber coils. It is treated by the addition and substraction of four semi-infinite coils of the type shown in (a).

* This work was assisted by the joint program of the ONR and AEC.

† Present address: Betatron Section, National Bureau of Standards, Washington, D. C.

[1] The calculations to be considered are concerned with circular coils of rectangular copper cross section. The results on rectangular coils of rectangular cross section can be found in the following papers: R. H. Lyddane and A. E. Ruark, Rev. Sci. Inst. **10**, 253 (1939); I. K. Beyerle, Archiv f. Elektrotechnik **25**, 269 (1931); and the experimental paper by E. Nageotte, Comptes Rendus **220**, 557 (1945).

[2] J. C. Maxwell, *Electricity and Magnetism* (The Clarendon Press, Oxford), third edition, (1904), Vol. II, p. 337.

[3] A. Gray, *Absolute Measurements in Electricity and Magnetism*, MacMillan Company, Ltd., London, second edition (1921).

[4] H. B. Dwight, *Electrical Coils and Conductors* (McGraw-Hill Book Company, Inc., New York, 1945).

[5] A. E. Ruark and M. F. Peters, J. Opt. Soc. Am. **13**, 205 (1926).

[6] M. Ference, A. E. Shaw, and R. J. Stephensen, Rev. Sci. Inst. **11**, 57 (1940).

[7] T. R. Lyle, Phil. Mag. **3**, 310 (1902).

[8] E. Von Zeipel, Arkiv. f. Mat. Astr. o. Fys. **30A**, N: 02 (1943).

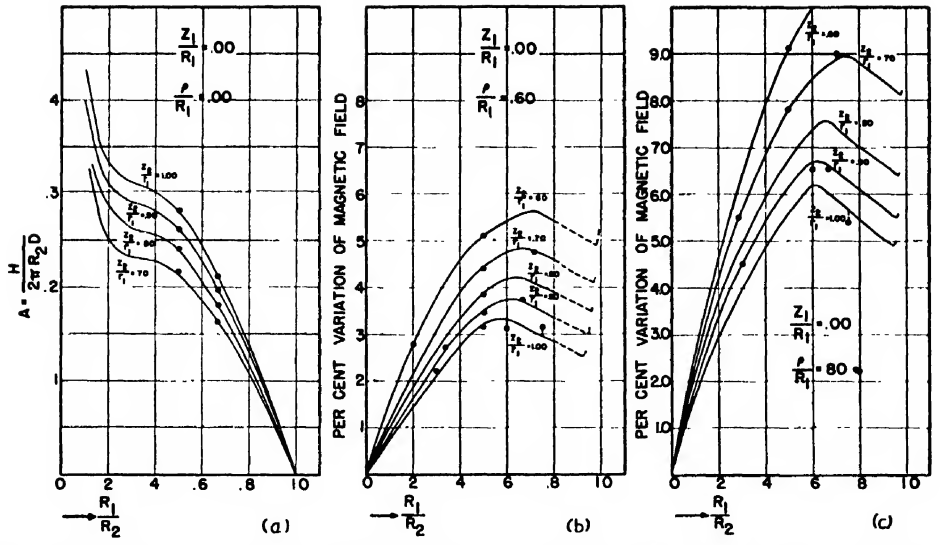


FIG. 2.

FIGS. 2–8. (a) Z component of the magnetic field intensity at $\rho=0$ in gauss-cm per abampere. (b) Z component of the magnetic field variation within a circle of radius ρ for $\rho/R_1=0.6$. The solid points represent calculated points as obtained from Table I. As can be seen, the curve shapes were only partially derived from the points, because of the inaccuracy associated with the points. The specific shapes were obtained from multiple plots and not from individual plots as presented here. The data obtainable from the (b) curves should be accurate to ± 0.1 percent. (c) Z component of the magnetic field variation within a circle of radius ρ for $\rho/R_1=0.8$. The data obtainable from the (c) curves should be accurate to ± 0.2 percent.

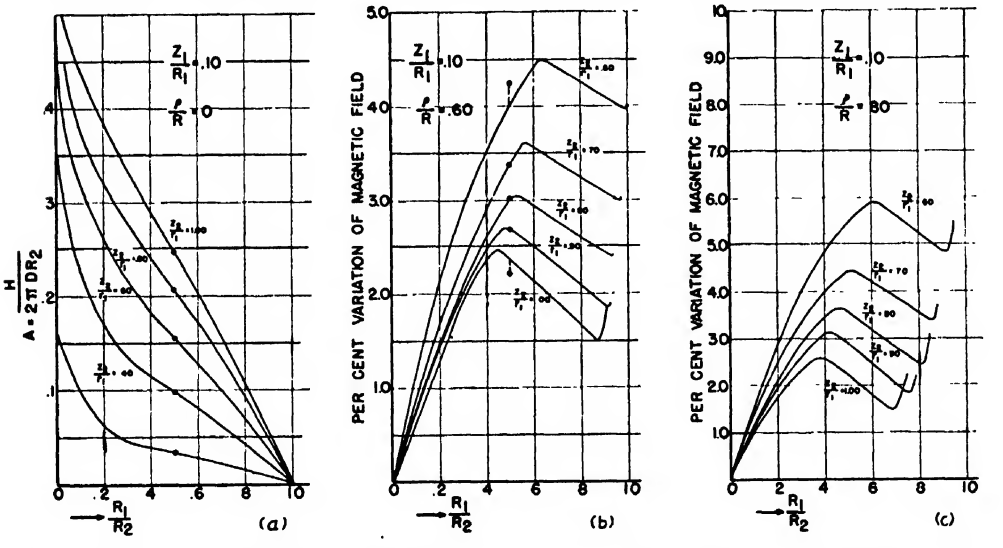


FIG. 3. (See Fig. 2 for caption.)

to the cylinders. Von Zeipel constructed coils to check the calculations.

Foss wrote two papers on the design of air core coils in which the integration procedure is used.[9]

For points close to the air core coils or for points between the coils, the calculations described have limited application because of the slow convergence of the series expressions for the magnetic field or because of the difficulty of integrating exactly over a desired cross section.

Also, calculations are not available on the magnetic field distribution of coils for which the ratio of the inner and outer diameters differ considerably from unity and for which the axial coil height becomes comparable to the radius.

This report will describe a different method for calculating magnetic field distributions, which makes it particularly adaptable to large cross-section coils. The method adopted is to find the field distribution due to a semi-infinite solid cylindrical coil of zero core diameter and unit winding density[10] as shown in Fig. 1(a). By the addition and subtraction of the field of four such semi-infinite coils the field distribution due to the finite hollow coil of Fig. 1(b) can be calculated. The results are still obtained from series expressions. However, the con-

[9] M. H. Foss, Carnegie Institute of Technology, ONR Contract, Technical Reports Nos. 2 and 3 (1948).

[10] The number of ampere turns per unit area of channel section will be called D.

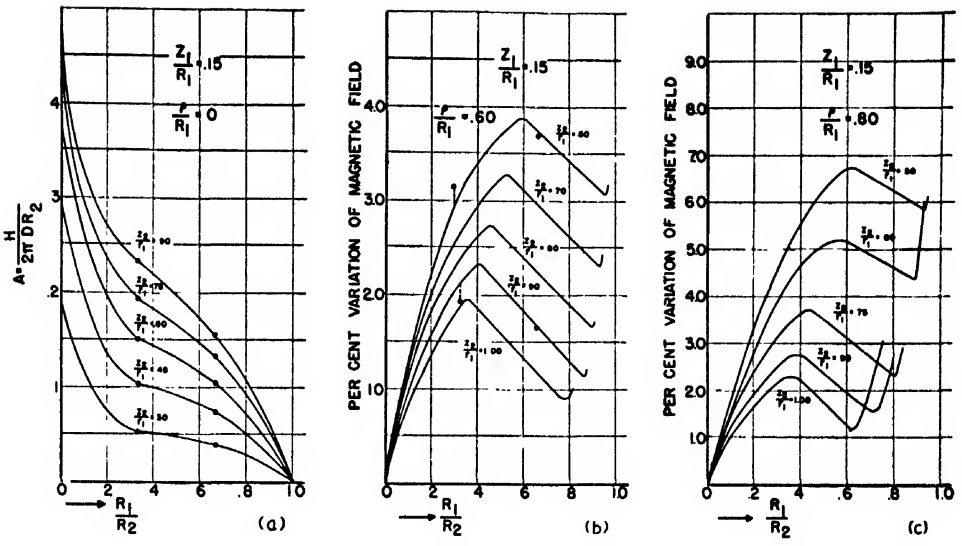


FIG. 4. (See Fig. 2 for caption.)

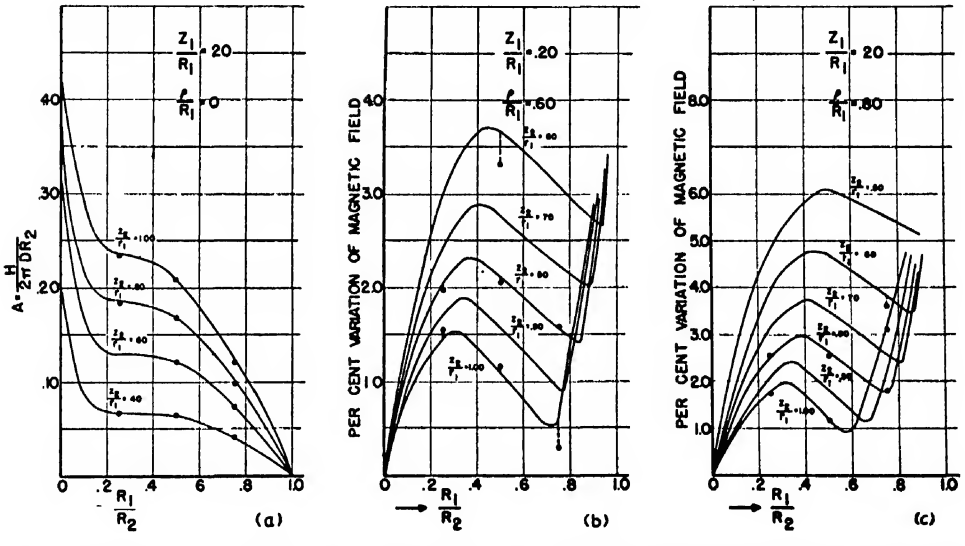


FIG. 5. (See Fig. 2 for caption.)

vergence is reasonable and permits, as an example, the non-uniformity of fields within a measuring circle ρ to be obtained to an accuracy of the order of ± 0.2 percent for ρ's as large as $\rho/R_1 = 0.8$.

The general technique to be described was first used by Butterworth[11] for calculating the mutual inductances for large cross-section coils. The procedure to be followed is very similar to that used by Cockroft[12] in the derivation of his equations. However, the present calculations give magnetic field distributions, while Cockroft used Butterworth's method to determine the mutual inductances and the expected mechanical forces in "Kapitza-type" coils.

THEORY

Consider a one turn coil of radius ρ at a distance of z from the end plane of the semi-infinite, solid coil of radius r, as in Fig. 1(a). The magnetic field in the z direction at any point on the circumference of the one turn coil will be

$$H_z = -\partial\Omega/\partial z, \tag{1}$$

where Ω is the magnetic scalar potential of the solid coil, which must satisfy Laplace's equation at all points external to the solid coil. For this case of circular symmetry about the z axis Laplace's equation can be written as

$$\frac{\partial^2\Omega}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial\Omega}{\partial\rho} + \frac{\partial^2\Omega}{\partial z^2} = 0. \tag{2}$$

A solution of Eq. (2) is

$$\Omega = \int_0^\infty A(\lambda) e^{-\lambda z} J_0(\lambda\rho)\, d\lambda, \tag{3}$$

where $J_0(\lambda\rho)$ is the zero-order Bessel function given by

$$J_0(\lambda\rho) = 1 - \frac{(\lambda\rho)^2}{2^2} + \frac{(\lambda\rho)^4}{2^2 4^2} - \frac{(\lambda\rho)^6}{2^2 4^2 6^2} + \cdots. \tag{4}$$

[11] S. Butterworth, Phil. Mag. 29, 578 (1915).
[12] J. D. Cockroft, Phil. Trans. Roy. Soc. 227A, 317 (1928).

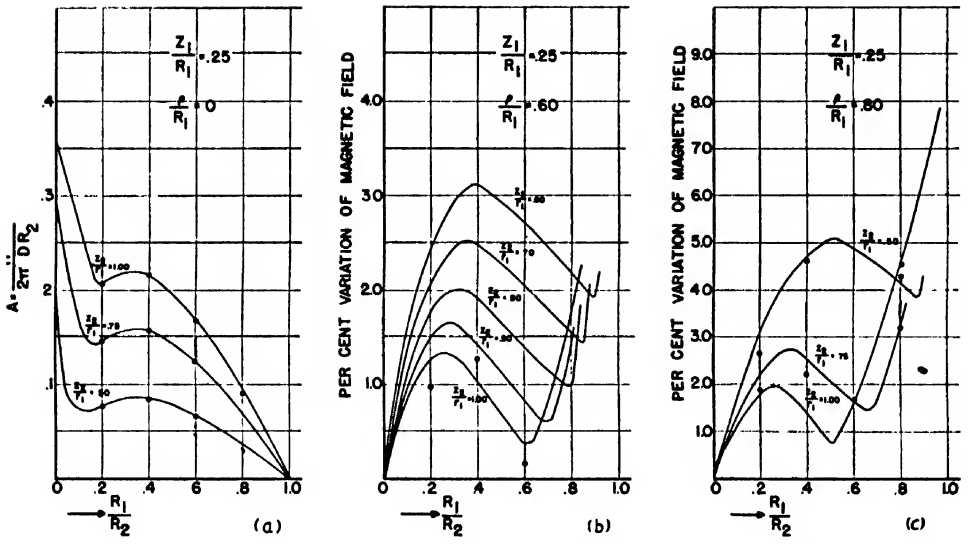


FIG. 6. (See Fig. 2 for caption.)

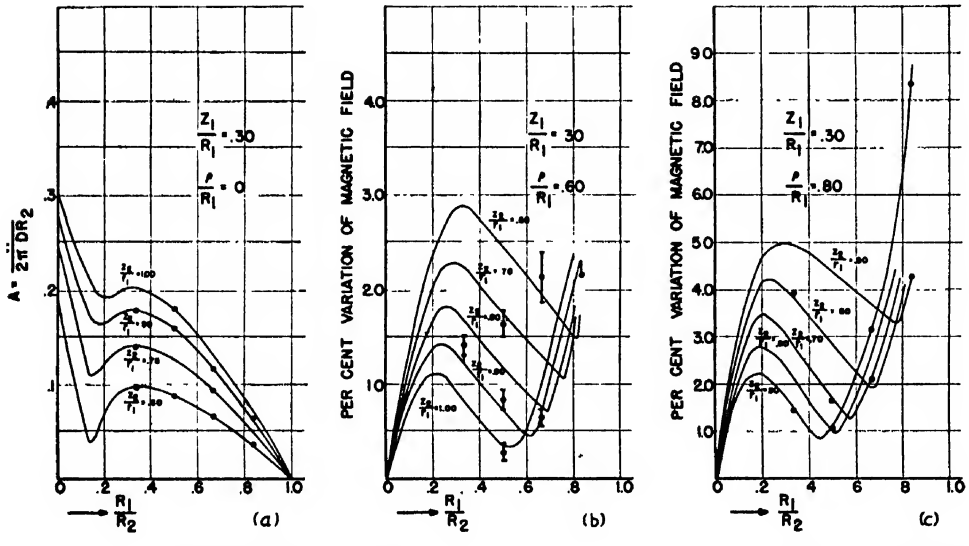


FIG. 7. (See Fig. 2 for caption.)

Case I—For $\rho/r \leq 1$

Since, when $\rho=0$, Eq. (3) reduces to

$$\Omega_0=\int_0^\infty A(\lambda)e^{-\lambda z}d\lambda, \tag{5}$$

Eq. (1) can be written as

$$H_z=\sum_{n=0}^{\infty}(-1)^{n+1}(\rho/2)^{2n}[1/n!]^2\partial^{2n+1}\Omega_0/\partial z^{2n+1}. \tag{6}$$

Ω_0 is the magnetic potential along the axis due to the semi-infinite coil of radius r and D abampere turns per centimeter squared. This is given by

$$\Omega_0=2\pi\int_z^\infty\int_0^r\left[1-\frac{z}{(y^2+z^2)^{\frac{1}{2}}}\right]Ddzdy \tag{7}$$

$$=\pi r^2D\left\{\frac{Z^2}{r^2}\log_e\left(\frac{1+\xi}{z/r}\right)+\xi-\frac{2z}{r}\right\}, \tag{8}$$

where $\xi=[1+(z^2/r^2)]^{\frac{1}{2}}$.

A. For $\rho/r \leq 1$ and $z/r > 1$

The integrand in (7) can be expanded for large z to give:

$$\frac{\Omega_0}{2\pi D}=\sum_{m=0}^{\infty}(-1)^m\frac{(2m+2)!}{[2^{m+1}(m+1)!]^2}\times\frac{1}{(2m+3)(2m+1)}\frac{r^{2m+3}}{z^{2m+1}}. \tag{9}$$

Therefore H_z becomes

$$\frac{H_z}{2\pi D}=\sum_{n=0}^{\infty}(-1)^{n+1}\left(\frac{\rho}{2}\right)^{2n}\left[\frac{1}{n!}\right]^2\times\frac{\partial^{2n+1}}{\partial Z^{2n+1}}\times\left\{\sum_{m=0}^{\infty}(-1)^m\frac{(2m+2)!}{[2^{m+1}(m+1)!]^2}\times\frac{r^{2m+3}/z^{2m+1}}{(2m+3)(2m+1)}\right\}. \tag{10}$$

B. For $\rho/r \leq 1$, $z/r < 1$, and $\rho/z < 1$

H_z from Eq. (6) is

$$H_z = -\frac{\partial \Omega_0}{\partial z} + \frac{\rho^2}{2^2}\frac{\partial^3 \Omega_0}{\partial z^3} - \frac{\rho^4}{2^2 2^4}\frac{\partial^5 \Omega_0}{\partial z^5} + \frac{\rho^6}{2^6 3^2 2^2}\frac{\partial^7 \Omega_0}{\partial z^7} - \frac{\rho^8}{2^8 4^2 3^2 2^2}\frac{\partial^9 \Omega_0}{\partial z^9} + \cdots \quad (11)$$

where the terms $\partial^{2n+1}\Omega_0/\partial z^{2n+1}$ can be determined directly from (8) for small z. It is convenient to separate Ω_0: $\Omega_0 = \Omega_1 + \Omega_2$ where

$$\Omega_1 = \pi D[z^2 \log_e\{r(1+\xi)\} + r^2\xi - 2zr]$$

and

$$\Omega_2 = -\pi D z^2 \log_e z. \quad (12)$$

A substitution of the derivatives of Ω_1 and Ω_2 into (11) results in:

$$\frac{H_z}{2\pi r D} = 1 + \frac{z}{r}\log_e\frac{z}{r} - \frac{z}{r}\log_e(1+\xi) - \left(\frac{\rho}{r}\right)^2 \times \frac{1}{4(z/r)\xi^3} - \left(\frac{\rho}{r}\right)^4 \times 1.5625 \times 10^{-2}\left\{\left(\frac{z}{r}\right)\left(\frac{3}{\xi^5} - \frac{15}{\xi^7} + \frac{1}{\xi^3} - \frac{1}{\xi(1+\xi)^2}\right)\right\} + \left(\frac{\rho}{r}\right)^6 \times 4.3403 \times 10^{-4} \times \left\{\frac{z}{r}\left(\frac{45}{\xi^7} - \frac{735}{\xi^9} + \frac{945}{\xi^{11}} + \frac{9}{\xi^5} + \frac{9}{\xi(1+\xi)^2} - \frac{6}{(1+\xi)^3} - \frac{3}{\xi^3}\right)\right\} - \left(\frac{\rho}{r}\right)^8 \times 2.0345 \times 10^{-5} \times \left\{\frac{z}{r}\left(\frac{525}{\xi^9} - \frac{18585}{\xi^{11}} + \frac{75}{\xi^7} - \frac{45045}{\xi^{15}} + \frac{58905}{\xi^{13}} - \frac{15}{\xi^5} + \frac{15}{\xi^3} - \frac{15}{\xi(1+\xi)^2} - \frac{30}{\xi(1+\xi)^3} - \frac{39}{\xi(1+\xi)^4}\right)\right\} + \left(\frac{\rho}{r}\right)^4 \times 3.1250 \times 10^{-2} \times \frac{1}{(z/r)^3} \times \left\{1 - 3.333 \times 10^{-1}\left(\frac{\rho}{z}\right)^2 + 1.563 \times 10^{-1}\left(\frac{\rho}{z}\right)^4 - 8.75 \times 10^{-2}\left(\frac{\rho}{z}\right)^6 + 5.469 \times 10^{-2}\left(\frac{\rho}{z}\right)^8 - 3.683 \times 10^{-2}\left(\frac{\rho}{z}\right)^{10} + 2.618 \times 10^{-2}\left(\frac{\rho}{z}\right)^{12} - 1.94 \times 10^{-2}\left(\frac{\rho}{z}\right)^{14} \cdots\right\}. \quad (13)$$

C. For $\rho/r \leq 1$, $z/r < 1$, and $\rho/z > 1$

For z less than ρ, the last bracketed term in (13) ceases to converge. This term is derived from the successive derivatives of Ω_2. Thus a distribution of poles must be found which gives the distribution of field

$$H_2 = \sum_{n=0}^{\infty}(-1)^{n+1}\left(\frac{\rho}{2}\right)^{2n}\left[\frac{1}{n!}\right]^2 \times \frac{\partial^{2n+1}}{\partial z^{2n+1}}(-\pi D z^2 \log_e z) \quad (14)$$

in regions for which $z < \rho$. To do this a method similar to that developed by Butterworth is used in which the magnetic field due to an axial distribution of poles extending from $z=0$ to $z=c$ with density $(-\pi z^2 D)$ is found. The field can be shown to be

$$\frac{H_2}{\pi D} = 3z + c - 2(\rho^2+z^2)^{\frac{1}{2}} - 2z\log_e\left[\frac{c+z+(\rho^2+(z+c)^2)^{\frac{1}{2}}}{c+z}\right] + 2z\log_e[z+(\rho^2+z^2)^{\frac{1}{2}}] + \frac{z^2}{(c+z)} - c - \frac{1}{\pi D}F(n, \rho, 1/c). \quad (15)$$

Since H_2 and Ω_2 are independent of the position of c, take $c=\infty$. For $c=\infty$, $F(n, \rho, 1/c)=0$ and

$$\frac{H_2}{\pi D} = 3z - 2(\rho^2+z^2)^{\frac{1}{2}} + 2z\log_e\left[\frac{z+(\rho^2+z^2)^{\frac{1}{2}}}{2}\right]. \quad (16)$$

Equation (16) can now be applied to give H_z for $z < \rho$. Equation (13) for this range of z's is replaced by

$$\frac{H_z}{2\pi r D} = 1 - \frac{z}{r}\log_e(1+\xi) + \frac{z}{r}\log_e\left[\frac{(z/r)+((\rho/r)^2+(z/r)^2)^{\frac{1}{2}}}{2}\right] - \left[\left(\frac{\rho}{r}\right)^2+\left(\frac{z}{r}\right)^2\right]^{\frac{1}{2}} + \left(\frac{\rho}{r}\right)^2 \times 0.25\left\{\frac{z}{r}\left(\frac{1}{\xi(1+\xi)} + \frac{1}{\xi^3}\right)\right\} - \left(\frac{\rho}{r}\right)^4 \times C + \left(\frac{\rho}{r}\right)^6 \times D - \left(\frac{\rho}{r}\right)^8 \times E, \quad (17)$$

where C, D, and E are terms which correspond to the ones in similar positions in Eq. (13).

TABLE I. $A=H_z/(2\pi Dr)$.

	$\rho/r=0$	$\rho/r=0.1$	$\rho/r=0.2$	$\rho/r=0.3$	$\rho/r=0.4$	$\rho/r=0.5$	$\rho/r=0.6$	$\rho/r=0.7$	$\rho/r=0.8$	$\rho/r=0.9$
$z/r=0.00$	1.00000000	0.90000000	0.80000000	0.70000000	0.60000000	0.50000000	0.40000000	0.30000000	0.20000000	0.10000000
$z/r=0.05$	0.81549483	0.7779395$_8$	0.7071310$_4$	0.6263370$_6$	0.541164$_8$	0.453777	0.3653$_4$	0.276$_7$	0.188$_7$	0.10$_2$
$z/r=0.10$	0.70017771	0.67795129	0.6262022$_9$	0.5607221$_3$	0.488339$_2$	0.41228$_9$	0.3345$_9$	0.256$_3$	0.179$_7$	0.10$_6$
$z/r=0.15$	0.61062320	0.59531780	0.55600348	0.5025254$_3$	0.4410210$_8$	0.37503$_1$	0.3069$_0$	0.2385$_6$	0.172$_6$	0.11$_0$
$z/r=0.20$	0.53751233	0.52609095	0.49522095	0.4511344$_9$	0.3988146$_7$	0.34166$_9$	0.2822$_0$	0.2225$_7$	0.165$_1$	0.11$_3$
$z/r=0.25$	0.47632186	0.46738471	0.44252487	0.40584188	0.3612455$_0$	0.31181$_3$	0.2600$_1$	0.2081$_0$	0.158$_4$	0.114$_0$
$z/r=0.30$	0.42433107	0.41712368	0.39672451	0.3659468$_4$	0.3278400$_7$	0.28508$_9$	0.2400$_3$	0.1948$_5$	0.151$_8$	0.114$_6$
$z/r=0.35$	0.37970336	0.37377216	0.35679817	0.33079058	0.2981425$_6$	0.261157	0.22196$_1$	0.1826$_2$	0.1453$_0$	0.112$_6$
$z/r=0.40$	0.34110756	0.33615670	0.32188356	0.29977423	0.2717273$_8$	0.23970$_0$	0.20559$_1$	0.1712$_7$	0.1387$_1$	0.110$_1$
$z/r=0.45$	0.30753199	0.30335642	0.29125683	0.27236551	0.2481706$_4$	0.220436$_4$	0.19072$_1$	0.160$_2$	0.1321$_7$	0.106$_9$
$z/r=0.50$	0.27818229	0.27463289	0.26431000	0.24809757	0.2272327$_9$	0.203114$_2$	0.17718$_9$	0.1509$_1$	0.1257$_6$	0.103$_2$
$z/r=0.55$	0.25241983	0.24933398	0.24053079	0.22656491	0.2084971$_9$	0.187511$_3$	0.16485$_4$	0.1417$_8$	0.1195$_5$	0.099$_3$
$z/r=0.60$	0.22972260	0.22711291	0.21948672	0.20741328	0.1917289$_7$	0.173432$_2$	0.15359$_2$	0.1332$_9$	0.1135$_7$	0.095$_4$
$z/r=0.65$	0.20965837	0.20740547	0.20081111	0.19034086	0.1766910$_3$	0.160704$_3$	0.14329$_4$	0.1253$_9$	0.1078$_7$	0.091$_6$
$z/r=0.70$	0.19186585	0.18991368	0.18419208	0.17508559	0.1631763$_6$	0.149178$_3$	0.13387$_2$	0.1180$_5$	0.1024$_4$	0.087$_7$
$z/r=0.75$	0.17604079	0.17434356	0.16936367	0.16142114	0.1510048$_4$	0.138720$_9$	0.12523$_1$	0.1112$_2$	0.0937$_0$	0.084$_0$
$z/r=0.80$	0.16192561	0.16044551	0.15609850	0.14915252	0.1400197$_4$	0.129214$_9$	0.11730$_4$	0.10487	0.0924$_4$	0.080$_4$
$z/r=0.85$	0.14930100	0.14800661	0.14420150	0.13811116	0.1300842$_4$	0.120558$_1$	0.11001$_8$	0.0989$_6$	0.0878$_4$	0.077$_0$
$z/r=0.90$	0.13797959	0.13684448	0.13350483	0.12815115	0.1210792$_7$	0.112502$_4$	0.10331$_8$	0.0934$_6$	0.0835$_0$	0.073$_7$
$z/r=0.95$	0.12780081	0.12680262	0.12386391	0.11914614	0.1129007$_8$	0.105438$_2$	0.09713$_9$	0.0883$_4$	0.0794$_0$	0.070$_6$
$z/r=1.00$	0.11862640	0.11774649	0.11515407	0.11098017	0.1054581$_8$	0.098841$_3$	0.09144$_2$	0.0835$_8$	0.0755$_4$	0.067$_8$

Case II—For $\rho/r>1$

Again the methods found in Cockroft's paper can be applied. The flux through a coil of radius ρ is

$$\phi_\rho=\int_0^\rho 2\pi\rho H_z d\rho=M_c I_c, \tag{18}$$

where M_c is the mutual inductance of the semi-infinite coil c in which the current I_c flows. Therefore,

$$H_z=(1/2\pi\rho)(\partial M_c/\partial\rho)\times I_c. \tag{19}$$

However,

$$M_c i_\rho=\int_0^r\int_z^\infty \phi_r D dr dz, \tag{20}$$

where

$$\phi_r=-\int_0^r 2\pi y\frac{\partial\Omega}{\partial z}dy=-\pi r^2\sum_{n=0}^{\infty}(-1)^n\left(\frac{r}{2}\right)^{2n}\times\frac{1}{n!(n+1)!}\frac{\partial^{2n+1}\Omega_0}{\partial z^{2n+1}}. \tag{21}$$

A substitution of (21) in (20) results in

$$M_c i_\rho=\pi D r^2\sum_{n=0}^{\infty}\frac{(-1)^n}{2^{2n}}\frac{r^{2n}}{2n+3}\frac{1}{n!(n+1)!}\frac{\partial^{2n}\Omega_0}{\partial z^{2n}}\Big|_z. \tag{22}$$

Therefore, from (19),

$$H_z=\frac{Dr^3}{2\rho}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^{2n}}\frac{r^{2n}}{2n+3}\frac{1}{n!(n+1)!}\frac{\partial^{2n}}{\partial z^{2n}}\left[\frac{\partial\Omega_0}{\partial\rho}\right] \tag{23}$$

where Ω_0 is the magnetic potential of a single turn along the axis and is given by

$$\Omega_0=2\pi\left\{1-\frac{z}{(\rho^2+z^2)^{\frac{1}{2}}}\right\}. \tag{24}$$

Since

$$-\frac{1}{2\pi\rho}\frac{\partial\Omega_0}{\partial\rho}=\frac{z}{(\rho^2+z^2)^{\frac{3}{2}}}, \tag{25}$$

H_z is given by

$$H_z=\pi D r^3\sum_{n=0}^{\infty}\frac{(-1)^n}{2^{2n}}\frac{r^{2n}}{2n+3}\frac{1}{n!(n+1)!}\times\frac{\partial^{2n}}{\partial z^{2n}}\left[\frac{z}{(\rho^2+z^2)^{\frac{3}{2}}}\right]. \tag{26}$$

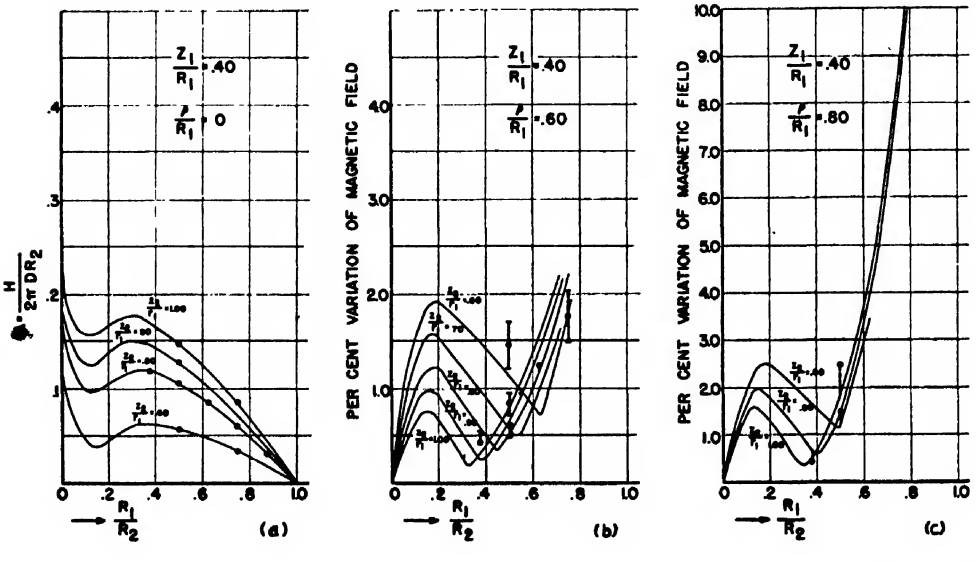


FIG. 8. (See Fig. 2 for caption.)

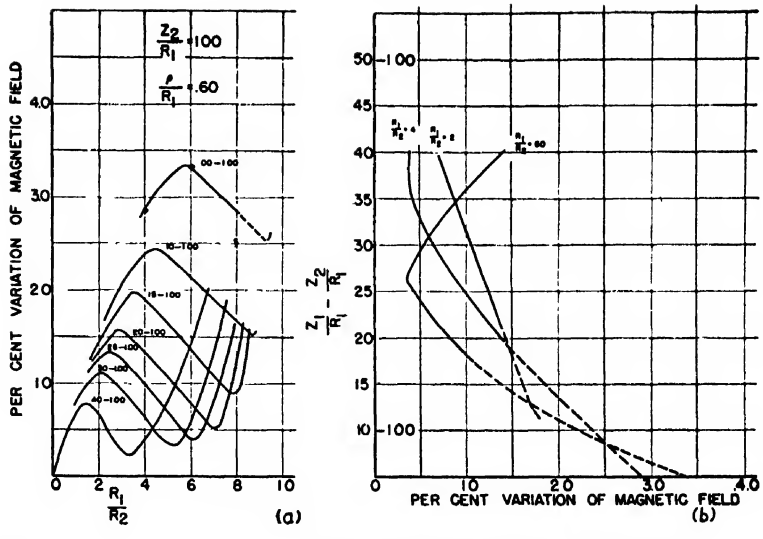


FIG. 9. (a) Percent variation of magnetic field z component variation *versus* R_1/R_2. The constants used for the curves are $\rho/R_1=0.6$ and $Z_2/R_1=1.00$. The variable is Z_1/R_1. The values of Z_1/R_1 and Z_2/R_1 are given above each curve. (b) The data in Fig. 9 (a) have been replotted with the values of Z_1/R_1 and Z_2/R_1 *versus* percent variation of magnetic field.

A. For $\rho/r>1$ and $z/r<1$

Take $\xi=(1+z^2/\rho^2)^{\frac{1}{2}}$. H_z can be evaluated by taking the successive derivatives indicated in (26). The result is

$$\frac{H_z}{\pi Dr^3}=\frac{1}{3}\frac{z}{\rho^3\xi^3}-\frac{3r^2z}{20\rho^6}\left[\frac{2}{\xi^5}-\frac{5}{\xi^7}\right]+\frac{r^4z}{336\rho^7}\left[\frac{120}{\xi^7}-\frac{840}{\xi^9}+\frac{945}{\xi^{11}}\right]-\frac{r^6z}{82{,}944\rho^9}\times\left[\frac{5040}{\xi^9}-\frac{68{,}040}{\xi^{11}}+\frac{187{,}110}{\xi^{13}}-\frac{135{,}135}{\xi^{15}}\right]+\frac{r^8z}{8{,}110{,}080\rho^{11}}\left[\frac{72{,}576}{\xi^{11}}-\frac{1{,}596{,}672}{\xi^{13}}+\frac{7{,}783{,}776}{\xi^{15}}-\frac{12{,}972{,}960}{\xi^{17}}+\frac{6{,}891{,}885}{\xi^{19}}\right]\cdots. \quad (27)$$

B. For $\rho/r>1$, $z/r>1$, and $\rho/z<1$

Ω_0 can be expanded in powers of ρ/z. H_z is then obtained by term by term differentiation of Eq. (26).

$$\frac{H_z}{2\pi rD}=-\frac{1}{6}\left(\frac{r}{z}\right)^2-\frac{1}{4}\left(\frac{r}{z}\right)^4\left(\frac{\rho}{r}\right)^2\left[1+\frac{3}{10}\left(\frac{r}{\rho}\right)^2\right]+\frac{5}{16}\left(\frac{r}{z}\right)^6\left(\frac{\rho}{r}\right)^4\left[1+\frac{6}{5}\left(\frac{r}{\rho}\right)^2+\frac{1}{7}\left(\frac{r}{\rho}\right)^4\right]-\frac{35}{64}\left(\frac{r}{z}\right)^8\left(\frac{\rho}{r}\right)^6\left[1+2.7\left(\frac{r}{\rho}\right)^2+\frac{9}{7}\left(\frac{r}{\rho}\right)^4+\frac{1}{12}\left(\frac{r}{\rho}\right)^6\right]+\frac{105}{256}\left(\frac{r}{z}\right)^{10}\left(\frac{\rho}{r}\right)^8\left[1+\frac{24}{5}\left(\frac{r}{\rho}\right)^2+\frac{36}{7}\left(\frac{r}{\rho}\right)^4+\frac{4}{3}\left(\frac{r}{\rho}\right)^6+\frac{3}{55}\left(\frac{r}{\rho}\right)^8\right]+\cdots. \quad (28)$$

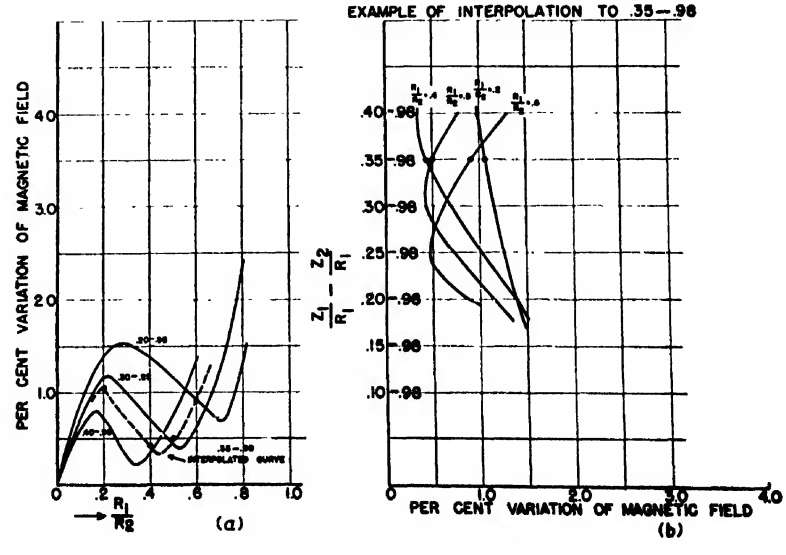


FIG. 10. (a) A plot of percent variation of magnetic field *versus* R_1/R_2 as obtained from the curves in Fig. 10(b). (b) An example of the interpolation to the intermediate values of $Z_1/R_1=0.35$ and $Z_2/R_1=0.98$. The data appearing in this graph were obtained from graphs of Z_1/R_1 and Z_2/R_1 *versus* percent variation of magnetic field [similar to Fig. 10(b)] as obtained from the data in Figs. 2(b), 3(b)···8(b).

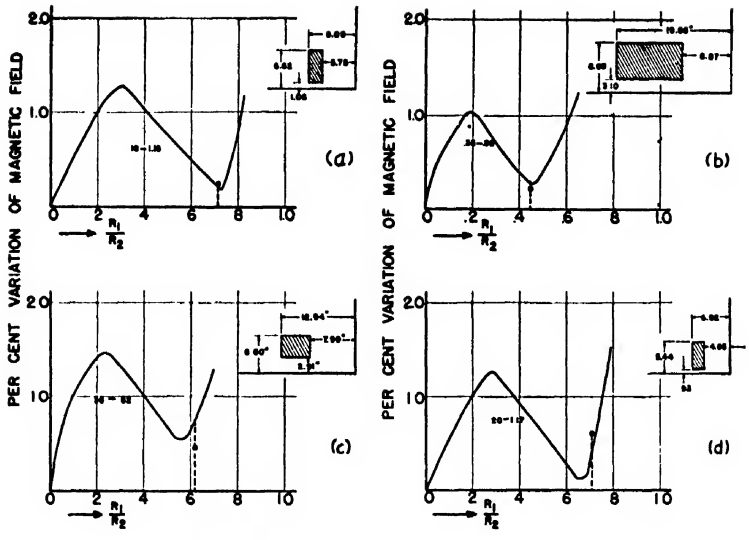


FIG. 11. (a) and (b) The data were obtained from two cloud-chamber coils built at the University of Illinois Betatron Laboratory. (c) The data were obtained from cloud-chamber coils built by Dr. J. R. Richardson in 1939 at the University of Illinois and from the field measurements of Dr. J. A. Phillips, which were privately communicated to the author. (d) The data were obtained from H. R. Crane (see reference 13).

RESULTS

Equations 10, 13, and 17 and Eqs. 27 and 28 permit the calculation of the z component of magnetic field, $H_z(r, z)$, due to a semi-infinite solid coil, whose radius is r and whose distance from the measuring plane is z. Therefore, the field $H_z(R_1, R_2, Z_1, Z_2)$ given by

$$H_z(R_1, R_2, Z_1, Z_2) = H_z(R_2, Z_1) - H_z(R_2, Z_2) - H_z(R_1, Z_1) + H_z(R_1, Z_2)$$

represents the field produced by one finite solid coil of the type shown in Fig. 1(b) and is the result of the addition and subtraction of the fields of four solid semi-infinite coils. The field in the z direction due to two, identical finite coils, one z_1 above the median plane and the other z_1 below the median plane, is twice that due to one coil. There is no radial component of field on the median plane with two coils, because the system is symmetrical.

The five equations enumerated, when the combination principle is used, allow the calculation of the field distributions both inside and outside the bounds of a pair of cloud-chamber coils. Since the present work was primarily concerned with the field distribution close to the axis of cloud-chamber coils and on the median plane, Eqs. 10, 13, and 17 were examined in regions for $\rho/r \leq 0.9$. Table I gives the field values, H_z, divided by $2\pi Dr$ for a semi-infinite solid coil at points for which $0 \leq z/r \leq 1.0$ and $0 \leq \rho/r \leq 0.9$. The values were all calculated using Eq. 13 and they have limited accuracies as indicated by the number of significant figures listed. Each value is in doubt by less than one in the last figure which in most cases has been subtyped.

The $A(=H_z/2\pi Dr)$ values in Table I were used to examine the absolute magnetic fields at $\rho/R_1=0$, 0.6, and 0.8 for a single finite coil whose dimensions are R_1, R_2, Z_1, and Z_2. In order to provide the results in a usable form, curves were drawn which give the fields on the coil axis [Figs. 2(a), 3(a), $\cdots$8(a)], the field variation within the circle of radius, ρ, for $\rho/R_1=0.6$ [Figs. 2(b), 3(b), $\cdots$8(b)], and the field variation for $\rho/R_1=0.8$ [Figs. 2(c), 3(c), $\cdots$8(c)]. The abscissa scale in each set of curves is R_1/R_2. In the case of the field variations within the circle of radius ρ, the percentage variation was obtained by dividing the difference between the maximum and minimum values of field within the circle of radius ρ by twice the mean value of field. All curves of the sets of three graphs (a, b, and c) apply to a constant Z_1/R_1 and a variable Z_2/R_1 as labeled.

The smooth curves, which are based on the calculated plotted points, should be used for predictions of field distributions. The smooth curves are more accurate than the calculated points, which are obtained from the data of Table I, because all the curves in Figs. 2–8 are internally consistent. This was accomplished by replotting the curves of Figs. 2–8 in inverse curves in which Z_2/R_1 was kept constant instead of Z_1/R_1. Any small inaccuracies in the calculated values were immediately apparent. An example of an inverse plot is shown in Fig. 9(a).

The calculated points of Figs. 7(b) and 8(b) were drawn with their estimated accuracies, in order to show the agreements of the points with the smooth curves.

The inverse plots of Figs. 9(a) and (b) are useful also for interpolation to a combination of z_1/R_1 and z_2/R_1, which is an uneven multiple of 0.05 and does not occur specifically in Figs. 2–8. As an example of interpolation, the field variation was desired for a z_1/R_1 equal to 0.35 and a Z_2/R_1 equal to 0.98. The interpolation was accomplished as shown in Figs. 10(a) and (b), which are inverse plots of the data in Figs. 2–8.

DISCUSSION

Little work has been done with Eqs. 27 and 28 because they deal specifically with points outside the coil bounds ($\rho/r > 1$). However, Eqs. 10, 13, and 17, which deal with points where $\rho/r < 1$, have been used extensively and several tests have been made to check their reliability. A crude test of Eq. 6, which forms the basis for Eqs. 10,

13, and 17, was made by differentiating with respect to r and z. The result is the formula for a circular current filament, which can be found in any standard book such as Gray's *Absolute Measurements in Electricity and Magnetism*.

A test was made of the values in Table I by calculating the first and second differences in the ρ/r and z/r directions. This served as a test of the numerical computations.

A test of the theoretically predicted values of field found in Table I and plotted in Figs. 2–9 was made by comparing the field variations predicted with those experimentally determined for four cloud-chamber coils.

The result of the test of the curves is shown in Figs. 11(a), (b), (c), and (d).

Figure 11(b) has drawn on it the cross section of one of the pair of coils to which the curve corresponds. As shown, the z_1/R_1 for the coil is 0.35 and the z_2/R_1 is 0.98. The curve in Fig. 11(b) shows the field non-uniformity for $\rho/R_1=0.6$ as a function of R_1/R_2. The procedure for interpolating to the intermediate values of 0.35 and 0.98 was given in Fig. 10(b). The plotted point for $R_1/R_2=0.45$, in Fig. 11(b) was determined experimentally by measuring the relative field values with two flip coils, each driven by $\frac{1}{4}$ H.P. synchronous motors. The experimental value so obtained is good to an accuracy of ± 0.05 percent. Since the theoretical curve should be accurate to ± 0.1 percent, the agreement of theory with experiment is good.

The only other comments to be made about Figs. 11(a), (c), and (d) are concerned with the comparisons of theory with experiment in the latter two figures. In Fig. 11(c) the experimental value is uncertain because of some uncertainty in the coil dimensions. These were obtained by rough measurements on the completed coil.

In Fig. 11(d) the dimensions were obtained from the drawing in the R.S.I. article by Crane[13] which was assumed drawn to scale.

The following procedure is a summary of the methods exemplified above for calculating the magnetic field characteristics for one annular coil whose dimensions are given by Z_1, Z_2, R_1, and R_2:

1. Calculate the ratios R_1/R_2, Z_1/R_1, and Z_2/R_1.

2. The z component of the magnetic field intensity at $\rho=0$ in gauss-cm per abampere (A) is given on the curves of Figs. 2(a), $\cdots$8(a) as a function of R_1/R_2, Z_1/R_1, and Z_2/R_1. For intermediate values which do not appear on the curves, use the interpolation procedure similar to that used in Fig. 10(a) and (b). D is in abamperes per cm^2 and A is equal to H_s for one coil divided by $2\pi DR_2$. The copper space factor of a non-ideal coil is never 100 percent and the correction factor must be included in the evaluation of the true current density, D.

3. The z component of the magnetic field variation within the circle of radius ρ is given as a function of the coil constants in Figs. 2(b) and (c), 3(b) and (c), $\cdots$8(b) and (c). It is believed that the values obtained from the (b) curves ($\rho/R_1=0.6$) are reliable to ± 0.1 percent and the values obtained from the (c) curves ($\rho/R_1=0.8$) are reliable to ± 0.2 percent. The curves should only be applied to coils, for which the ratio of R_1/R_2 is between 0.1 and 0.9.

The tedious numerical computations involved to obtain Table I and the figures in this paper were capably performed by Mrs. Nelson Grisamore and Mr. Norton Moise. The author acknowledges the helpful criticism of Dr. E. F. Shrader, who carefully reviewed this manuscript.

[13] H. R. Crane, Rev. Sci. Inst. 8, 440 (1937).

The Effect of Compressive Stresses on the Linear Thermal Expansion of Magnesium and Steel

JOSEPH L. ROSENHOLTZ AND DUDLEY T. SMITH
Department of Geology and Mineralogy, Rensselaer Polytechnic Institute, Troy, New York
(Received October 19, 1949)

A series of cylinders of annealed, extruded, pure magnesium and medium soft steel were subjected to a sequence of compressive stresses at room temperature and their respective coefficients of linear thermal expansion in the direction of stress application were determined for the range 20° to 100°C. The cylinders were then heat treated to permit either recovery or recrystallization and the changes in length were measured. Characteristic curves are given for each metal in which the coefficient of linear thermal expansion and length changes are plotted as a function of the true compressive stress. It is shown that these properties may be used to determine the stress history of a test specimen for which characteristic curves are available. It is suggested that this stress analysis technique may find application in the study of rock deformation and in the investigation of metals, both before and after failure.

INTRODUCTION

IN a comparative study of a single crystal of the Iceland spar variety of calcite and Yule marble, in which approximately half of the calcite grains have their optic axes in preferred alignment, it was found[1] that the volume thermal expansion of the marble was twice as great as for the calcite. Furthermore, the calcite showed no change in length after cooling to room temperature from 700°C, while the recovery in the marble resulted in an increase in length in excess of 1 percent. Both phenomena are clearly associated with the deformational stresses of unknown magnitude which produced the marble and which were relieved during the heat treatment. This explanation received substantiation when it was found that a second heating-cooling cycle gave coefficients of expansion of the same order of magnitude as for the unstressed single crystal of calcite.

Before beginning a study of the effects of high, confining, compressive stresses on a series of marble specimens, it seemed desirable to investigate metals in order to ascertain what type of expansion changes, if any, might be expected. Toward this end, extruded, pure magnesium was selected for study because several of its physical and mechanical properties are somewhat similar to those of calcite. Among these are anisotropy, large linear coefficient of thermal expansion, α, and deformation by both twinning and slip. The second metal selected was a medium soft steel containing 0.29 percent carbon and 0.43 percent manganese. While this was nominally a hot rolled steel, metallographic examination gave evidence of preferred grain orientation indicating that the rolling was finished somewhat below the critical point. This was desirable from the point of view of this research because it permitted the study of a two-phase system in which sufficient deformation had occurred to produce some grain distortion in the direction of rolling, which was also the direction of compression and testing.

[1] J. L. Rosenholtz and D. T. Smith, Am. Mineral 34, 846 (1949).

PROCEDURE

Test specimens of each metal were cut from single bars and were machined to $\frac{3}{8}$ in. diameter and $\frac{3}{4}$ in. long. The magnesium cylinders were annealed for 1 hour at 300°C; the steel cylinders received no heat treatment because it was desired to retain the original structure. The study of a preliminary series of magnesium specimens showed very significant changes which indicated the particular stress regions where more numerous determinations were required. A complete magnesium series was then subjected to progressively greater compressive stresses to the point of rupture at 19 T.S.I. The maximum stress used for the steel corresponded to the region where failure began by extensive plastic deformation in the vicinity of 40 T.S.I.; no fracture occurred up to 135 T.S.I. As a practical expedient, all specimens were allowed to stand at room temperature for at least two weeks to permit recovery to take place.

A specially designed quartz dilatometer having a sensitivity of 2–3 microinches was used for determining α. The equipment, which has been fully described,[1] consists of a transformer bridge and includes a recorder which gives a complete record of dilatation.

Temperatures were measured with a Type K2 Leeds & Northrup potentiometer using a Pt-Pt 10 percent Rh thermocouple. The furnace was precooled to 10°C before measurements were begun and the heating rate was controlled so that the 20° and 100° points were accurately correlated with expansion. In all cases, the values of α are for the range 20° to 100°C.

As will be shown presently, specimens subjected to widely different compressive stresses may have the same value of α. As a result, a unique solution of stress cannot be obtained from an expansion-stress curve alone. Length changes resulting from recovery or recrystallization appeared to give a second useful relationship. The length recovery of magnesium cylinders after heating to 100°C was too erratic to be helpful. However, annealing at 300°C for one hour gave results consistent with the changes in α. Since the steel cylinders showed no change in length after cooling from 100°C, they were

heated in vacuum for one hour at 650°C for relief of stresses. After cooling to room temperature they were again heated in vacuum to 850°C, the length of the cylinders being measured before and after each heat treatment. A consistent set of values was found as a result of recovery but the results of the full anneal were erratic and could not be used.

STEEL

The results of the investigation of the relationship between α and compressive stress of medium soft steel is shown in Fig. 1. True stresses are used since they reveal the actual compression per unit of area for the respective applied stresses.

Smith and Wood[2] determined the stress-strain curve for the atomic lattice of mild steel (0.1 percent C) in compression, their measurements having been made perpendicular to the direction of stress application. Their curve displays some of the same characteristics as Fig. 1 with the principal exception that it reaches a maximum and then remains constant soon after the yield point has been reached. They divided the behavior of the lattice in terms of three phases. In the first phase, in which the compressive stresses are below the yield point, the crystal lattice obeys Hooke's law and recovers completely when the load is removed. The second phase appears at the yield point where an abrupt, permanent, lattice contraction amounting to 0.03 percent was found. In the third phase, which occurs beyond the yield point, they suggest a "regular increase with stress of the permanent residual lattice strain and the superposition of this residual strain on the reversible elastic expansion induced by the applied stress. The two types of strain are so related that their algebraic sum results in an approximately constant lattice spacing at stresses beyond the yield."

Several important characteristics which did not appear in the strain of the lattice of the mild steel referred to above are found in Fig. 1. In the first place, *AB* corresponds with the first phase of Smith and Wood. However, α does not remain constant as it should have if all stresses were relieved when the compressive loading was removed. While the percentage increase in α is small, it nevertheless increases up to the yield point. It is suggested that this effect is produced by the plastic deformation of some of the steel crystals while most of them are reacting elastically. This seems to correspond to the second stage of deformation of polycrystalline aggregates as described by Boas.[3] The residual stresses surrounding those crystals which experienced plastic deformation may therefore account for the increase in α which becomes progressively larger as the yield point is approached.

The second phase is represented by *BC* which is at the yield point. Here an abrupt drop in α was found in excess of 3 percent. This obviously corresponds to the sudden permanent contraction in the lattice. The third phase, corresponding to *CD*, and amounting to only 6 percent of the total stress range of the curve, gave a phenomenally large increase in α and coincides with Smith and Wood's region above the yield point where the lattice expanded.

The most important departure from agreement with permanent lattice changes is found along *DE*, where Smith and Wood obtain a constant value but where α decreases as much as it previously increased. If we assume, as suggested, that a constant lattice spacing is due to the algebraic sum of the permanent lattice expansion and the reversible lattice expansion, then the marked decrease in α must be some measure of the extensive plastic deformation produced by the relatively high compressive stresses. It appears, therefore, that there is a threshold value, represented by the point *D*, where the spacing decreases while the residual stresses become correspondingly greater. This suggestion is consistent with the explanation of the process of recrystallization given by Seitz[4] who explains that the energy stored during cold-work presumably furnishes the driving force for the recrystallization.

There were no changes in the length of the steel cylinders after cooling from 100°C. However, the changes after heating to 650°C for one hour were quite unexpected since, as shown in Fig. 2, there was either no change or a contraction. At the beginning of this investigation, the dimensional changes were observed as a matter of record. However, when it was found that the steel might possess the same value of α for several widely different compressive stresses, a related property was sought for stress correlation whereby a unique solution would be obtainable. This solution, then, is found for the steel in the length changes induced by a stress relief anneal. The inflection in the length-stress curve coincides with the maximum of the expansion-stress curve.

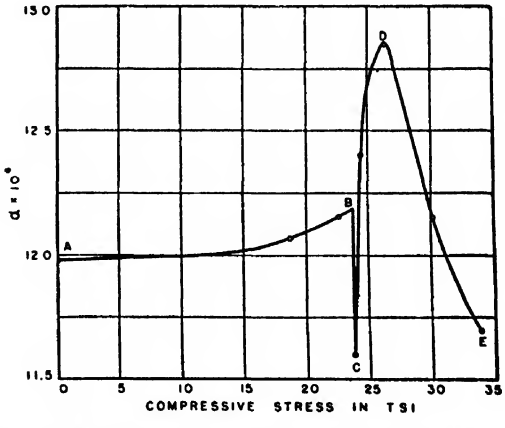


FIG. 1. The relationship between coefficient of linear thermal expansion (20°–100°C) and true compressive stress for medium soft steel.

[2] S. L. Smith and W. A. Wood, Proc. Roy. Soc. 181, 72 (1942).
[3] W. Boas, *An Introduction to the Physics of Metals and Alloys* (John Wiley and Sons, Inc. New York, 1947), p. 103.
[4] F. Seitz, *The Physics of Metals* (McGraw-Hill Book Company, Inc., New York, 1943), p. 81.

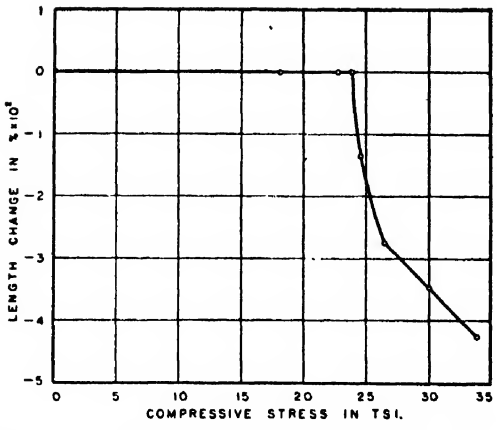


FIG. 2. Percentage change in length of compressed, medium soft steel cylinders after heating for one hour at 650°C.

As for the mechanism of contraction, the abrupt contraction of the lattice spacing at the yield point referred to above coincides with the increase in diameter when plastic deformation begins throughout the steel. As a result, the residual stresses appear to be stored in such a fashion that, when recovery takes place by the heat treatment, expansion of the lattice occurs perpendicular to the direction of compression and thereby causes an actual length decrement.

MAGNESIUM

The effect of compressive stresses on thermal expansion for pure, extruded, annealed magnesium is shown in Fig. 3. The characteristics of the curve are similar to those for steel except for the portion below the yield point. All changes in α at and above the yield point are approximately twice as great as those for steel, but α for magnesium is more than double that of steel. The gradual decrease in α below a true compressive stress of 3 T.S.I. is caused by the fact that measurable permanent deformation occurred, as contrasted with no permanent deformation in steel under a corresponding stress. In other words, appreciable plastic deformation began when the smallest stress was applied and it increased gradually until the abrupt change occurred at the yield point. While it is frequently said that magnesium has no definite yield point, the experimental evidence presented above leads to the opposite conclusion.

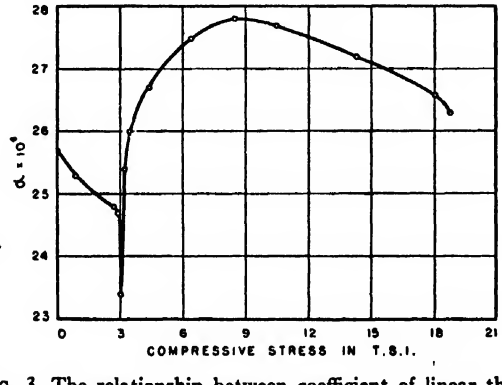


FIG. 3. The relationship between coefficient of linear thermal expansion (20°–100°C) and true compressive stress for pure, extruded, annealed magnesium.

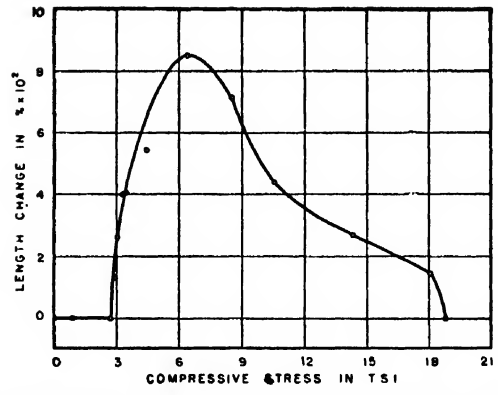


FIG. 4. Percentage change in length of compressed magnesium cylinders after annealing for one hour at 300°C.

As in the case of steel, the same value of α might correspond to two stress values. Measurements after cooling from 100°C could not be used because it was obvious that recovery or recrystallization or both had taken place. Recrystallization may occur in magnesium at 100°C, particularly if it has been stressed. Inasmuch as the different cylinders were stressed to varying degrees, the length changes were too inconsistent to be of use in selecting the correct stress corresponding to a particular value of α.

The results after annealing for one hour at 300°C are shown in Fig. 4. No length changes were measurable up to the yield point, after which the curve is quite similar to Fig. 3. The explanation proposed for the behavior of steel applies equally well to the magnesium. It is possible to arrive at a unique solution of stress from both of these curves except near the peak of the expansion curve where a close approximation would be obtainable.

SUMMARY

It is obvious from the experimental results presented that the coefficient of linear thermal expansion of a material is subject to wide variation and depends upon its previous thermal and deformational history. It is very important to be aware of the fact that if an anisotropic single crystal possesses widely different directional values of α, a polycrystalline specimen of the same material may show an amazingly great variation in α, especially if preferred orientation is pronounced. Disregarding instrumental methods and all attendant factors which pertain to precision of measurement, the published values of α for many materials are so widely divergent as to indicate the importance of knowing their previous history.

The method described above can lead to at least a close approximation of the stresses imposed upon metals

which experience plastic deformation. It is of the utmost importance, however, that the material used in the determination of the characteristic expansion-stress and length-stress curves have the same history as a test specimen for which the unknown superposed stress history is to be determined. This method of stress analysis may be especially useful in the analysis of failure in metals and perhaps in the limit design of structures.

As stated in the introduction, this research was undertaken in an attempt to anticipate how brittle minerals or rocks might react if subjected to compressive stresses under confining pressures. Griggs[5] has already demonstrated that some rocks, such as limestone and marble, behave like metals under high confining pressures. They display an elastic limit and deform by plastic flow when stressed above this point. It is anticipated, therefore, that some minerals and rocks which have been subjected to high, differential pressures will give characteristic curves somewhat similar to those of steel and magnesium. If this proves to be the case, a method will be available for achieving a better understanding of the range of pressures which have been responsible for the great deformational effects produced during metamorphism.

[5] D. Griggs, J. Geol. 44, 541 (1936).

ACKNOWLEDGMENT

The authors are grateful for the assistance provided by the Rensselaer Polytechnic Institute Research Fund. They also express their appreciation for the helpfulness of Professor Joseph F. Throop of the Department of Mechanics and Mr. Stephen E. Wiberley of the Department of Chemistry of Rensselaer Polytechnic Institute.

Thermal Noise at High Frequencies

A. VAN DER ZIEL*

Department of Physics, University of British Columbia, Vancouver, Canada

(Received September 6, 1949)

Spenke's discussion of thermal noise from the point of view of the electron theory leads to a result which differs from Nyquist's original formula at frequencies ν such that $h\nu/kT \geqq 1$. This discrepancy is due to the assumptions which Spenke made in his analysis.

1. INTRODUCTION

ABOUT 20 years ago Nyquist[1] published his well-known thermodynamical discussion of thermal noise; since then it has been shown that Nyquist's results could be derived with the help of the electron theory of conduction (Bernamont,[2] Bakker and Heller,[3] and Spenke[4].) Unfortunately Spenke's and Nyquist's results differ at frequencies ν such that $h\nu/kT \geqq 1$. It will be shown that the source of this discrepancy is in Spenke's treatment of the problem.

For our discussion we have to define "available power." A signal generator has an e.m.f. e and an internal impedance $Z=R+jX$; its available power P_a is defined as the maximum power which can be dissipated by an external impedance Z'. This maximum occurs if $Z'=R-jX$, so that:

$$P_a=\tfrac{1}{4}e^2/R. \tag{1}$$

2. THERMODYNAMICAL THEORY OF THERMAL NOISE

Nyquist's[1] analysis consists of three steps:

(a) The available noise power P_a in a small frequency interval $\Delta\nu$ of *any* thermal noise generator is a *universal* function of the frequency ν and the absolute temperature T of the generator. He proved this with the help of the second law of thermodynamics.

(b) The available noise power P_a in a frequency interval $\Delta\nu$ around a central frequency ν is:

$$P_a=\bar{E}\Delta\nu, \tag{2}$$

where $\bar{E}$ is the average energy of a harmonic oscillator of frequency ν and absolute temperature T (a system with two degrees of freedom). Nyquist proved this in a somewhat artificial way but his result cannot be doubted. A more detailed proof of (2) was given by Schremp.[5]

(c) $\bar{E}$ is given by the equipartition law. The average energy of an harmonic oscillator of frequency ν and temperature T is:

$$\bar{E}=kT\cdot f(\nu);\quad f(\nu)=h\nu/kT[\exp(h\nu/kT)-1]^{-1}, \tag{3}$$

and one obtains:

$$P_a=kT\Delta\nu\cdot f(\nu). \tag{4}$$

$f(\nu)=1$ if $h\nu/kT \ll 1$; at those frequencies:

$$P_a=kT\Delta\nu. \tag{5}$$

Equation (5) has been verified experimentally; the deviation between (4) and (5) becomes important in the

* Now at: Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota.

[1] H. Nyquist, Phys. Rev. 32, 110 (1928).
[2] J. Bernamont, Ann. d. Physik. Paris 7, 71 (1937).
[3] C. J. Bakker and G. Heller, Physica 6, 262 (1939).
[4] E. Spenke, Wiss. Veroeff. Siemenswerke 18, 54 (1939).

[5] E. J. Schremp, in *Valley and Wallman*, "Vacuum Tube Amplifiers," M.I.T. series (McGraw-Hill Book Company, Inc., New York, 1948), Vol. 18, p. 529.

infra-red (at least at normal room temperature) and no experimental verification of (4) has been published up to now.*

In order to prove (4) in a less artificial way consider a hollow sphere at a uniform temperature T containing an harmonic oscillator of frequency ν_0 and a circuit consisting of a self-induction L, a capacity C and a resistance R and being tuned at the frequency ν_0. The hollow sphere is filled with black-body radiation of temperature T; the harmonic oscillator and the tuned circuit both interact with this radiation and will assume the temperature T. Their average energy E will be the same and the equipartition law will hold. Half of the average energy of the circuit is stored in the condenser, hence:

$$\tfrac{1}{2}C\langle v^2\rangle=\tfrac{1}{2}\bar{E}, \tag{6}$$

where v is the fluctuating voltage across C. A Fourier analysis shows that P_a is proportional to $\Delta\nu$:

$$P_a=\alpha\Delta\nu. \tag{7}$$

We now have to prove: $\alpha=\bar{E}$. Let the thermal noise of the circuit for the frequency interval $\Delta\nu$ be described by an e.m.f. e in series with R, then after (1):

$$\langle e^2\rangle=4P_aR=4\alpha R\Delta\nu. \tag{7a}$$

The contribution of e to the mean square value of the noise voltage v across C in the frequency interval $\Delta\nu$ is:

$$\langle e^2\rangle\left|1+\left(\frac{1}{j\omega L}+j\omega C\right)R\right|^{-2}=4\alpha R[1+4(\omega-\omega_0)^2C^2R^2]^{-1},$$

if $(\omega-\omega_0)\ll\omega_0$ and $\omega_0^2LC=1$. Integrating over all intervals $\Delta\nu$:

$$\langle v^2\rangle=\alpha/C \quad\text{or}\quad \tfrac{1}{2}C\langle v^2\rangle=\tfrac{1}{2}\alpha \quad\text{or}\quad \alpha=\bar{E}.$$

3. THE CORPUSCULAR THEORY OF THERMAL NOISE

The following model of a conductor is used here: Electrons in the conductor move in all possible directions, they have a Fermi velocity distribution at a temperature T. The electrons which have a velocity close to the top of the Fermi distribution can gain energy when an electric field is applied and thus give rise to conduction. Those electrons can collide with the positive ions of the lattice after having traveled an average "free path length" Ω; as they all have practically the same velocity v, the average time of flight τ_0 along the free path length Ω is Ω/v. It is assumed that the collisions with the positive ions of the lattice occur without an exchange of energy and that the electrons are scattered in random directions after such a collision.

The motion of electons along the free path length Ω gives rise to current pulses of length τ_0 in the conductor (thermal noise). Carrying out a Fourier analysis of the resultant noise current $I(t)$, Bakker and Heller[3] and also Spenke[4] found for its Fourier components i:

$$\langle i^2\rangle=\langle i_0^2\rangle(1+\omega^2\tau_0^2)^{-1}, \tag{10}$$

where $\langle i_0^2\rangle$ is the l.f. value of $\langle i^2\rangle$.

If a d.c. field is applied to the conductor, the electrons gain energy in the field; due to the collisions with the lattice a state of equilibrium is established in which the electrons have an average velocity in the direction of the field. This drift velocity, which depends upon τ_0, constitutes a current and so the collisions with the lattice give rise to a finite d.c. conductivity G_0. If an a.c. field is applied, the conductivity G remains independent of frequency as long as its period is small in comparison to τ_0. If it is comparable to τ_0, the electrons will on the average gain less energy in the field, as its phase changes appreciably during the motion along the free path length; at those frequencies $G<G_0$. Spenke[4] found:

$$G=G_0(1+\omega^2\tau_0^2)^{-1}; \tag{11}$$

this result was first obtained by Kronig.[6] Hence:

$$P_a\equiv\langle i^2\rangle/4G\equiv\langle i_0^2\rangle/4G_0. \tag{12}$$

Evaluation of $\langle i_0^2\rangle$ and G_0 then yields (5). This holds for all frequencies.

Spenke's result is not a very surprising one, for in his model no exchange of energy occurs when the electrons collide with the lattice. The average energy stored in the condenser C of an LCR circuit tuned at the frequency ν_0 will then be in equilibrium with the thermal agitation of the electrons. The average energy for translatory motion as given by the equipartition law does not contain the Planck factor $f(\nu)$ and hence one would expect that, independent of the frequency ν_0:

$$\tfrac{1}{2}C\langle v^2\rangle=\tfrac{1}{2}kT$$

so that (5) should be valid for all frequencies. Nowhere in this model is any room left for an expression containing $h\nu/kT$ and this cannot be improved by using another velocity distribution either.

But in a genuine thermodynamical discussion it is not sufficient to state that the electrons have a velocity distribution corresponding to the temperature T; one has to immerse the conductor in a heat reservoir of large heat capacity and one has to investigate how the spontaneous current fluctuations in the conductor are coupled to this heat reservoir. Undoubtedly this coupling occurs as follows: The heat reservoir is in thermal equilibrium with the lattice vibrations and the electrons couple the current fluctuations to those vibrations. This coupling is due to the fact that in each collision an energy exchange $\pm h\nu$ occurs between the colliding electron and the lattice (ν is one of the frequencies of the lattice vibrations). The average energy stored in the condenser C of an LCR circuit tuned at the frequency ν_0

* It would be worthwhile to test its validity in the centimeter wave-length region at low temperatures. If $T=1.2°K$ then $h\nu/kT=1$ at 1 cm wave-length ($\nu=3\times10^{10}$); the deviation between (4) and (5) is then about 40 percent, which might be detectable with present techniques.

[6] R. de L. Kronig, Proc. Roy. Soc. A124, 409 (1929); 133, 255 (1931).

will therefore be in equilibrium with the average energy $\bar{E}$ of the *lattice vibrations* of frequency ν_0, and $\bar{E}$ is equal to the average energy of an harmonic oscillator given by (3). Hence in the correct theory:

$$\tfrac{1}{2}C\langle v^2\rangle=\tfrac{1}{2}\bar{E}$$

so that we obtain (4) instead of (5).

4. THE SOURCE OF THE ERROR

As it is now clear that Spenke's analysis is wrong, it has to be investigated whether the error is in (10) or in (11). According to the Fourier analysis, fluctuating quantities $X(t)$ have Fourier components x_n of frequency ω_n such that:

$$\langle x_n^2\rangle=4\Delta\nu\int_0^\infty\langle X(t)X(t+w)\rangle\cos\omega_n w dw. \quad (13)$$

The correlation function $\langle X(t)X(t+w)\rangle$ is zero if $|w|\ll\tau$, where τ is the *correlation time* of the fluctuating quantity. If $(\omega_n\tau)\ll 1$:

$$\langle x_n^2\rangle=4\Delta\nu\int_0^\infty\langle X(t)X(t+w)\rangle dw.$$

This integral is independent of the frequency and is completely determined by the equipartition law; any theory using this law will give the right formula at low frequencies even though it gives wrong values for τ and for $\langle X(t)X(t+w)\rangle$. In the latter case the noise spectrum will have the wrong shape at high frequencies.

Lameris[7] found from his reflection measurements in the infra-red that (11) was correct. He deduced from his measurements that at room temperature $\omega\tau_0$ was equal to unity at the frequencies $\nu(=\omega/2\pi)$: 2.36×10^{13} for Ag, 2.39×10^{13} for Au and 5.23×10^{13} for Pt. But after (4) the available noise power shows a marked decrease if $h\nu/kT=1$, this corresponds to $\nu=0.6\times10^{13}$ at room temperature. Hence the correlation time τ of the fluctuations is about 5–10 times as large as the correlation time τ_0 of the resistance. But in Spenke's model $\tau=\tau_0$ so that the error is definitely in (10).

Spenke[4] made the following assumptions:

(a) After collisions the electrons are scattered in random directions. In fact the scattering over small angles is favored.

(b) No energy is exchanged in the collision process. In fact each collision corresponds to an exchange of energy with the lattice. It is therefore likely that an exact calculation of $\langle i^2\rangle$, in which the above facts are taken into account will give (4) instead of (5). The calculation is very difficult, whereas the thermodynamical analysis is very simple.

In order to show that the two above facts work in the right direction, consider a number of electrons dN having a velocity between v and $v+dv$. In calculating the correlation time of the resistance we have the problem: "What is the probability that the electrons will obtain a different velocity v' when traveling along the distance dx?" This will give a correlation time equal to the time τ_0 between two collisions, if on the average the electrons lose the energy gained by traveling in the electric field after each collision (for a more detailed discussion compare Spenke's[4] paper). In calculating the correlation time τ of the current fluctuations the problem is: "How long does it take before there is no longer any correlation between the current due to the motion of these dN electrons and the original current?" If the scattering over small angles is favored, it will take a number of collisions before this is achieved so that τ is much longer than τ_0, just as was required.

5. AVAILABLE NOISE POWER OF AN ANTENNA

Burgess[8] calculated the available noise power P_a in the frequency interval $\Delta\nu$ of an antenna receiving black body radiation of uniform temperature and showed that it was given by (4). That this leads to (4) and not to (5) can be shown by a simple argument.

At low frequencies the energy density of black body radiation is:

$$U_\nu\equiv(8\pi\nu^2/c^3)\cdot kT\Delta\nu \text{ (Rayleigh-Jeans' law).} \quad (14)$$

But at those frequencies $P_a=kT\Delta\nu$, hence:

$$U_\nu=(8\pi\nu^2/c^3)\cdot P_a. \quad (15)$$

Equation (15) describes a property of the antenna which does not depend upon quantum phenomena so that it will hold for all frequencies. But at high frequencies (14) has to be replaced by Planck's law. As this means that (14) has to be multiplied by the Planck factor $f(\nu)$, P_a has to be multiplied by the same factor, so that (4) holds.

The author is indebted to Dr. A. J. Dekker and Dr. F. A. Kaempffer for many discussions on this subject.

[7] P. F. Lameris, Ph.D. Thesis, Groningen, 1936.

[8] R. Burgess, Proc. Phys. Soc. 53, 293 (1941).

Semiconductive Colloidal Suspensions with Non-Linear Properties

HANS E. HOLLMANN
U. S. Naval Air Missile Test Center, Pt. Mugu, California
(Received October 3, 1949)

Semiconductive particles in a colloidal suspension, when subjected to an electric polarizing field, form semiconductive chains. Once the suspension is polarized, its conductivity increases with increasing fields because the forces of electrostatic attraction press the individual particles together, thus strengthening the transition contacts along the chains. Both effects result in an over-all electric non-linearity that is the subject of extensive investigation.

The non-linear characteristics of such colloidal resistors are oscillographed and described in relation to the polarizing field strength by means of simple formulas based on field coefficients of the first and second order.

In addition to a true rectification, resulting from the curved characteristics, a "pseudo-rectification" occurs which corresponds to an average pre-polarization by stronger alternating fields. The alternating forces of electrostatic attraction assure a high frequency response, so that non-linear Lissajous-figures and harmonics, up to radiofrequencies, are produced.

ACCORDING to Debye's theory of polar molecules the behavior of certain substances, especially of polar liquids and polar dielectrics, such as crystals and ceramics with their abnormally high dielectric constant and permittivity, their rotational and frictional dispersion, their Curie temperatures, etc., is caused by the polarization and orientation of their dipole molecules or their molecule-complexes, when subjected to an electric polarizing field.

For the following considerations it is important to recall that the microcrystalline dipoles originally have a random orientation but, when polarized, orient themselves in the direction of the electric lines of force.

In addition to these polar fluids and substances comprised of natural dipole-molecules, it is possible to enter a more macroscopical region in which the molecules or molecule-complexes are replaced by bigger (but nevertheless microscopically small) particles suspended in a liquid—and later embedded in a solid carrier, preferably in colloidal dispersion. In this way, the concept of polar substances is to be transposed to that of polar suspensions. Attention is drawn to these types of polar liquids by the so-called "magnetic fluid"[1] which represents the prototype in the magnetic field just as ferromagnetism relates to ferroelectricity. For the sake of completeness and easier understanding, it may be convenient to recall the characteristics of the magnetic fluid before we turn to the new polar suspensions in the purely electric field.

The sprinkling of iron filings on a surface under the influence of a magnetic field is a well-known method for making magnetic lines of force visible. All the iron particles become magnetized, attract each other, and arrange themselves in the form of chains in accordance with the magnetic lines of force.

This old phenomenon has been put to ingenious use in the magnetic fluid. Instead of sprinkling onto a surface through free air, the iron particles, or more accurately fine iron powder, is mixed with oil, thus forming an iron-oil suspension, wherein the iron particles are much more mobile than on a solid surface. As soon as the magnetic fluid is subjected to a magnetizing field, it thickens, and if magnetized sufficiently becomes of a paste-like consistency. Hence, the magnetic oil is a new type of a thixotropic fluid with the important difference, that its gel-sol-gel transformation can be controlled by means of an external magnetic field instead of by shearing or stirring.

The solidifying effect can be demonstrated by dipping a permanent magnet into the suspension, after which a certain region around the poles becomes so thickened that it can be pulled out in the form of a semisolid bridge.

The magnetic fluid has found an important application in the magnetic fluid clutch. Instead of utilizing the immediate forces of attraction occurring between the two coupling elements of a mechanical or pure magnetic clutch, a small gap between the coupling elements is filled with the magnetic fluid. Under the influence of a magnetic control field, the magnetic fluid becomes a tough semisolid, binds the coupling elements together, and transmits the torque from the driving to the driven element.[2]

The magnetic fluid, furthermore, may be exploited extensively in electromagnetically controlled hydraulic systems where the controllable thixotropic fluid itself replaces the conventional valves.[3] Another application is for the purpose of forming casting molds where the fluid, in which the model has been immersed, solidifys instantaneously and produces a workable mold as long as it is subjected to a strong magnetic field. Finally, the magnetic fluid may be utilized for forming a controllable resistor similar to a relay. However, such an oil-relay

[1] Jacob Rabinow, "Magnetic fluid clutch," Tech. News Bull. Bur. Stand. 32, 54–60 (May, 1948).

[2] G. R. Nelson, "Magnetic fluid clutch in servo applications," Electronics 100–103 (November, 1949).

[3] "Magnetic fluid uses," Electronics 120–122 (September, 1949). "New uses for magnetic fluids," J. Frank. Inst. 248, 155–157 (1949).

will not de-energize by itself but, in order to restore its initial resistance must be stirred or shaken.

Turning to the field of electrostatics, we obtain the so-called "electric fluid."[4] It can be best understood by recalling the so-called "Lichtenberg figures" which in the field of electrostatics are identical to the aforementioned magnetic figures. A light powder having a high dielectric constant, e.g., lycopodium (club moss), in an electrostatic field, forms chains similar to those of the iron filings in the magnetic field, thereby making visible the electric lines of force. Consequently, intermixing or suspending a fine dielectric powder in an insulating liquid of lower dielectric constant—the simplest being in oil—produces the electric fluid whose viscosity also increases transversely to the applied electric field. Hence, it is possible to apply the aforementioned magnetic experiment to the field of electrostatics. Two suitable electrodes or wires, across which either a high direct or alternating voltage is applied, are dipped into the electric fluid. When the wires are withdrawn, a little bridge or membrane adheres. Naturally, as soon as the polarizing voltage is cut off the dielectric suspension assumes its neutral viscosity and drops off.

It is clear that the electrically controlled thixotropy is much weaker than the magnetic chain effect. This is because the electrostatic forces of attraction are limited basically by the breakdown field strength. This disadvantage, however, is compensated for by a very low consumption of control power. Consequently, an electric clutch, of which the torque moment is transferred via the thickened electric fluid, may be of some practical value in cases where a high frequency response and low control power is required.

Fundamentally, the electric fluid and its thixotropy is not completely new. The oldest example is blood. The analogy was observed years ago, in the field of high frequency diathermy.[5] In that part of the body subjected to the diathermy field the blood particles become polarized and also form dielectric chains. Without going into further detail, it may be mentioned that the increased viscosity hampers the blood circulation in the treated vessels, thus counteracting the increased circulation in the vessels which have been enlarged physiologically under the influence of the diathermic heating; hence, the dielectric chains are in constant conflict with the blood circulation, fortunately without endangering the regulative heat flow.

Another characteristic of the electric fluid is the change of its dielectric constant in relation to the applied electric field. A capacitor containing the electric fluid as dielectricum becomes non-linear, i.e., its capacity is a function of the impressed voltage. In this respect, the electric fluid resembles certain polar substances, i.e., rochelle salt, similar crystals, and ceramics with Seignette-electric properties. Referred to the viscosity effect it is obvious that the electric non-linearity occurs long before any mechanical effect can be observed. It would be beyond the scope of this paper to go into greater detail concerning this electric characteristic.

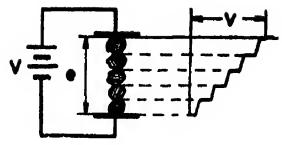


FIG. 1. Qualitative representation of a semiconductive chain and the incremental voltage drop.

After the foregoing introduction, particularly concerning the non-linear capacitor, we proceed to a new concept, namely, an "electric oil" having non-linear or controllable conductivity. For this purpose the dielectric particles must have a certain conductivity which will enable the chain formation to increase the over-all conductivity at the expense of the viscosity effect which now loses its significance. Finally, the electric oil can be mixed or composed in a manner so that its conductivity exceeds its dielectric constant by several orders of magnitude. As a result the wattless component vanishes at the expense of the dissipative component. It is therefore possible to make a non-linear resistor whose value depends on the applied electric field as well as on various external or internal influences. As will be shown in the following pages, the electric oil displays a surprising versatility and creates new fields for numerous applications.

1. THE FORMATION OF SEMICONDUCTIVE CHAINS

In the course of numerous experiments with various mixtures and suspensions, it was found that a suspension of graphite in oil produces very satisfactory results. Besides relatively crude mixtures of coarse graphite powder and lampblack in silicon oil, some colloidal suspensions, such as Oildag, Castordag, and Glydag, operate very stably, particularly so because no sedimentation takes place. Although it is probable that these compositions do not represent an optimum, and taking into consideration that further investigation may disclose better samples, the present investigations are concerned only with graphite-oil suspensions by means of which the unusual behavior and versatility of the electric oil can be proven very satisfactorily.

The process of chain formation can be observed under a microscope. In neutral condition or in the virgin state the particles are distributed irregularly. As soon as the suspension is subjected to a polarizing field, the particles form chains. The picture, however, is not ideal because the primary chains, extending mostly in the direction of the electric field lines, are crisscrossed by secondary

[4] W. M. Winslow, U. S. Patent 2,417,850 (March, 1947). "Induced fibration of suspensions," J. App. Phys. 20, 1137–1140 (1949).

[5] W. Krasny-Ergen, "Nicht-thermische Wirkungen elektrischer Schwingungen auf Kolloide" (Non-thermal effects of electric oscillations on colloids), Zeits. f. Hochfrequenzt. u. Elektroakustik 48, 126–133 (1936); Report of the International Congress for UHF-Waves in Physics, Biology, and Medicine, Vienna (1937), pp. 180–184.

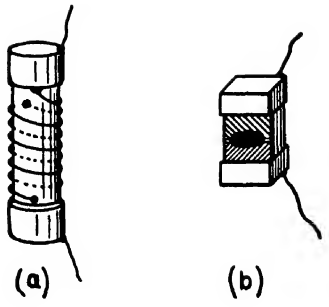


FIG. 2. Two examples of colloidal resistors.

chains extending between points of different potentials along the primary chains, as a result of which a complex network appears instead of individual parallel chains.

The chain formation alone, however, does not suffice to explain the non-linear conductivity. Once the chains are formed at a certain field, a second effect is responsible for the non-linear property, namely, the fluctuating forces of electrostatic attraction along every individual chain, or more accurately, between contacting particles or links. This is explained in greater detail by means of the qualitative representation shown in Fig. 1. The left side illustrates a simple chain between two electrodes with the separation e and the electric field $E=V/e$ where V is the voltage. The right side illustrates roughly the voltage drop along the chain. The semiconductivity of the particles, together with the high transition resistance of the regions where the particles touch each other, effect an incremental voltage drop. Partial fields appear preponderantly at the transition regions, whereas the conductivity prevents higher voltage drops across the particles themselves.

Furthermore, the relatively high fields created in the transition regions produce strong electrostatic forces of attraction between all adjacent particles. As a result, the dielectric chains are reinforced at increasing field strengths so that the increasing over-all conductivity of every individual chain, and therefore of the sum of all chains, may be considered as being caused by the action of innumerable fluctuating contacts between innumerable links. Once the chains are formed, the conductivity continues to increase with increasing field strength, i.e., the over-all conductivity of the electric oil becomes non-linear. A very important result of this contact action is the fact that the non-linearity does not depend on any movement of the particles through the carrier oil but only upon the reinforced pressure along the chains which produces the non-linear conductivity and assures a high frequency response. (A similar effect of strong electrostatic attraction at moderate voltages is known as the Johnson-Rahbek effect[6] occurring in the microscopically small airgap between the surface of a semiconductor, e.g., slate or agate, and an impressed foil electrode.)

[6] K. Rottgardt, "Elektrische Anziehung nach Johnson-Rahbek und ihre Anwendung" (Electrostatic attraction according to Johnson-Rahbek and its application), Zeits. f. Tech. Physik 2, 315 (1921).

The ratio between the dielectric constant of the particles and that of the carrier oil, the conductivity of the particles, and their size, is also of great importance. It can be seen at once that the foregoing theory leads to an important conclusion. If the individual chain contains, for example, N links, the partial voltage drops between subsequent links can be considered as V/N if the potential drop across the particles themselves, in rough approximation, is neglected. Consequently, noticeable forces of attraction occur only under the assumption that V/N exceeds a certain minimum value. This simple relationship determines the applicability of the electric oil, or, more accurately, its non-linearity and voltage sensitivity. The latter quantities can be chosen by matching the electrode separation to the operating voltages, e.g., by selecting the suitable full operating field strength in relation to the size of the particles.

2. COLLOIDAL RESISTORS

For any practical applications or experimental investigations, the space between suitably arranged electrodes must be filled with electric oil so that a non-linear colloidal resistor is obtained. In analogy to a fluid capacitor, a simple form would consist of two flat or cylindrical electrodes with a fixed or variable separation arranged in an insulated container filled with the oil. If an electrical stability as high as possible is desired, care must be taken that there is no surplus amount of oil outside of the electrodes because the electrid field gradually attracts the dielectric particles from the surroundings so that the mixture ratio in the

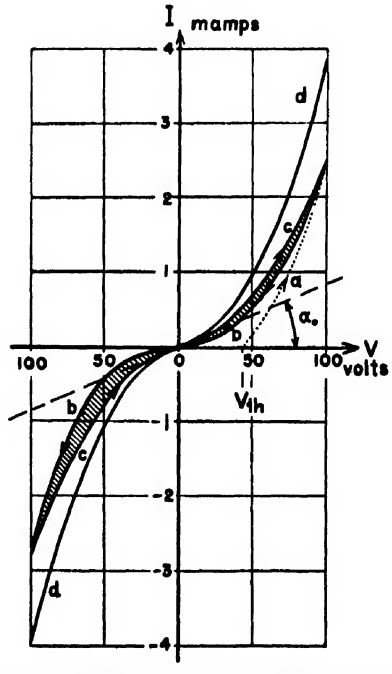


FIG. 3. I–V characteristic of a colloidal resistor at various states of polarization.

inter-electrode space increases at the expense of the surplus oil. Another modification is a glass tube closed at both ends and filled with the oil in which wire electrodes are immersed at each side.

Two more convenient types of oil-resistors are schematically indicated in Fig. 2. Type (a) contains two bare wires which are bifilarly wound around an insulating rod. The non-linearity or voltage sensitivity depends on the separation of the wires, in other words, on the pitch, whereas the resistance range or more accurately, the zero resistance R_0 at very low voltages is related to the surface of the electrodes, i.e., to the length of the bifilar helix.

When a non-linearity as high as possible is required, the second type (b) is very convenient. A minute electrode gap is made by cutting a slit into the silver layer of a little mirror, thus assuring a high polarizing field even at moderate voltages. Naturally, this mirror type may be combined with the former form by cutting a helical slit into a metal-plated rod or by cutting a meandering line similar to a Rowland grating.

Instead of being immersed in an oil-filled container, the electrode-gaps may be covered with a thin layer of electric oil which may be protected by wrapping or hermetically sealing. Whether one type or the other is preferred depends on the various conditions of operation. It must be clearly understood that every colloidal resistor with a larger volume of free oil is more or less microphonic, whereas the sensitivity to vibration vanishes in all cases where only a thin layer of the oil is supported by an insulator.

The possibility of matching the power dissipation as well as the voltage sensitivity to any desired range is of great practical significance. Whereas the power dissipation is determined largely by the cross section or by the current density—if necessary with air or water cooling —the voltage sensitivity or non-linearity, respectively, depends for a given suspension only on the oil gap. Accordingly, the two electrode arrangements shown in Fig. 2 allow encompassing a wide range of non-linearity as well as resistivity.

In addition to the preceding considerations, the colloidal resistor may be compared to the previously mentioned magnetically controlled resistor because both devices have in common the fluctuating transition resistances between individual particles. An important difference, however, is that the electric oil leads to a steadily increasing transition or controllability whereas the magnetic fluid relay operates, so to speak, only in "black-and-white," i.e., its resistance fluctuates only between infinity and zero. Furthermore, the electric oil requires no artificial reduction to its initial state, in contrast to the fluid relay. Only when the electric oil is overloaded do its particles burn and fuse together, and then they must be broken up by stirring, or by strong vibrations.

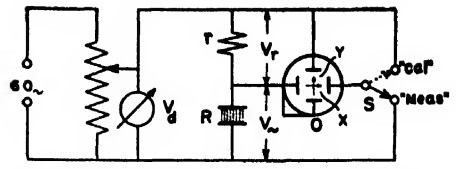


FIG. 4. Simple device for oscillographing non-linear I–V characteristics.

3. NON-LINEAR CHARACTERISTICS

(a) Static Measurements

For obtaining a preliminary insight into the non-linear properties of colloidal resistors, numerous types were investigated at static conditions. In these experiments, the resistors were merely fed by a variable direct voltage V and the passing currents I were observed. Some typical examples of simple I-V characteristics are shown in Fig. 3 in the form:

$$I=f(V)=V/R=G\cdot V, \qquad (1a)$$

where R and G are the non-linear resistance and the non-linear conductance, respectively.

A detailed study of the characteristics yields the following results: At the beginning of the measurements, the electric oil is in its virgin state. Because no chains are present, current cannot flow. As soon as the impressed voltage reaches a certain threshold value V_{th}, the chains become active and produce a semiconductive connection between the electrodes. From then on the current rises rapidly with the voltage and follows the steep virgin curve (a) thereby revealing the progressive strengthening of the chains by the increasing forces of electrostatic attraction. When the chains, under the influence of a certain polarization, have attained a certain state, they keep this assumed state, at least to a certain degree, even though the voltage is decreased. In this particular example not enough time has elapsed at the extreme maximum voltage to permit the chains to become fully developed. Consequently, the first I–V characteristics (b) and (c) do not coincide but circumscribe the shaded area.

In order to describe the static picture as completely as possible, it must be emphasized that only a relatively small range can be measured. As noted before, the electric oil becomes unstable at higher current densities because the graphite particles burn and fuse together so that the original colloidal character is completely destroyed.

Quite obviously, the phenomenon of the chain formation has a certain analogy to hysteresis, with the important difference being that the chain formation has a certain inertia. As a result, the chain hysteresis disappears, even at low frequencies, because the chains assume an average state of polarization.

In regard to this time lag and, therefore, to the previous history, the static step-by-step method is somewhat delicate and laborious. A steady state is

reached only after a sufficient time has elapsed before the next step. Accordingly, only after a sufficiently strong polarization has taken place, preferably connection with some reversals of the polarizing field, the fairly stable I–V characteristic (d) in Fig. 3 was measured.

This final characteristic may be considered in greater detail. It passes the zero point of inflection according to the dashed tangent with the slope $\tan\alpha_0$ indicating a zero resistance R_0 or a zero conductance G_0. At higher fields the electrostatic attraction causes the currents to rise over this initial tangent thus indicating a decrease in resistance or an increase in conductivity. This preliminary analysis picture will be completed later.

(b) Oscillograms

Because of the aforementioned phenomena, the static step-by-step method is somewhat inaccurate. An oscillographic representation is much more advantageous because it discloses not only the dynamic behavior of the electric oil in question, but, at the same time, assures greater stability because the chains assume an average state of polarization.

The simplest device for oscillographing non-linear I–V characteristics is indicated in Fig. 4. The colloidal resistor R is fed via the fixed preresistor r by the alternating voltage V_d of 60 cycles with variable amplitude. The cathode ray of the oscilloscope O—which, for the sake of simplicity, is shown without its amplifiers—is deflected horizontally by the voltage V_- across R, if the switch S is in "measuring" position, and vertically by the voltage drop $V_r = I \cdot r$.

Figure 5 shows a typical example of a non-linear I–V oscillogram (a). The straight calibration line (b) was obtained with the switch S in "calibration" position so that both inputs of the oscilloscope are in parallel. If Y and X are the deflecting sensitivities in

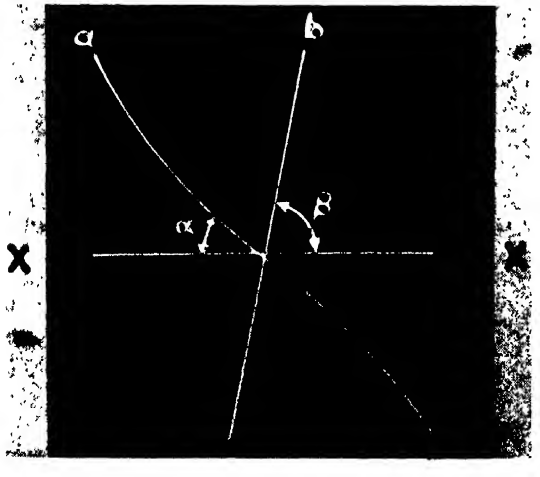


Fig. 5. Oscillographic I–V characteristic (a) with calibration line (b).

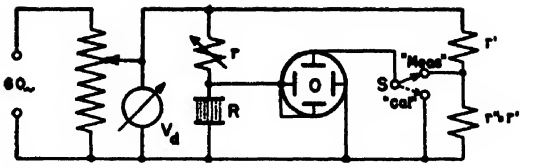


Fig. 6. Alternating current bridge for obtaining oscillographic bridge characteristics.

both axes, then their ratio is determined by the angle β: $Y/X = \tan\beta$. On the other hand, any angle α of an arbitrary chord connecting the point of any voltage V with the zero point or two arbitrary points at any voltages $\pm V$, whose tangent is proportional to the corresponding conductance $G = 1/R$, is given by the ratio of the horizontal deflection:

$$D_x = X \cdot V_d/(1+r/R)$$

and the vertical:

$$D_y = Y \cdot V_d/(1+R/r)$$

as:

$$\tan\alpha = D_y/D_x = (r/R)\tan\beta = r \cdot G \cdot \tan\beta. \qquad (2)$$

By means of this relation, R, G, or $\sigma = e/F \cdot R$ (F = surface of the electrodes) can be plotted *versus* V provided the horizontal deflection sensitivity (X) is determined by a simple voltage calibration.

A more sensitive device for oscillographing the non-linearity of a colloidal resistor is an alternating current bridge whose schematic is indicated in Fig. 6. It differs from Fig. 4 only in that the non-linear attenuator $r \cdots R$ is supplemented by the resistive attenuator $r' \cdots r''$, which balances the fundamental voltage drop across R so that only the non-linear deviations produce the vertical deflection.

A simple analysis as before yields:

$$\tan\alpha = \tan\beta \frac{(r'r/r''R)-1}{(r'/r'')+1}. \qquad (3a)$$

Naturally the bridge has its highest sensitivity when $r' = r''$ so that Eq. (3a) simplifies into:

$$\tan\alpha = \tfrac{1}{2}\tan\beta[(r/R)-1]. \qquad (3b)$$

The bridge operates at its maximum sensitivity if it is in balance at very low momentary voltages. This means that r must always be adjusted in such a manner that the middle portion of the oscillogram remains horizontal. Such an example is shown in Fig. 7 together with the calibration line (b) which is obtained as before with the switch "S" in "calibration" position.

In analogy to Fig. 3 the gradual formation of the chains can be demonstrated by increasing the alternating driving voltage V_d. Such an oscillogram is shown in Fig. 8 whose individual characteristics differ by driving voltages progressing in increments of five volts. As long as the colloidal resistor is in its virgin state, R is approximately infinite and a little deflection line

appears with the slope:

$$\tan\gamma \sim \tfrac{1}{2}\tan\beta.$$

As soon as the threshold field is reached, the oscillogram becomes non-linear and, after having passed a horizontal middle portion at $R=r$, simultaneously becomes erect and expands.

This turning, especially the middle portion, reveals a decrease of the zero resistance R_0. On the other hand, the bridge can easily be kept in balance by always adjusting r in such a manner that the turning-up is neutralized or that the non-linear bridge characteristic keeps a horizontal zero passage. The measuring points shown in Fig. 9 were obtained in this manner and will be explained in greater detail in Section 5.

(c) Analytical Characteristics

A complete analysis of the complex behavior of the colloidal resistors is not possible, at least not at the present time, because too many parameters are unknown and not computable. However, an analytical representation of the experimental I–V characteristics yields a better insight into the philosophy of the electric oil and electrical performance of any colloidal suspension. Further investigations may disclose the complex relations by which the individual parameters, i.e., the dielectric constant and conductivity of the particles and the carrier fluid, the form and size of the particles and their surface properties, the mixture ratio, and last but not least, the previous history are related.

The following analysis is based on numerous measurements of a graphite-oil suspension. The solid curve in Fig. 10 shows an example which has been averaged from numerous measuring points at a certain prepolarization and limited to a permissible current density. The characteristic is deduced by plotting the current density i per cm^2 instead of the entire current I, and the field strengths E, instead of the impressed measuring voltage V. Hence, Eq. (1a) is presented by the reduced form:

$$i=I/F=\sigma E=\sigma V/e \qquad (1b)$$

where F is the electrode area in cm^2, σ the conductivity in mhos per cm, and E the electric field strength in volts per cm. This reduction has an advantage in that the characteristic is referred directly to the conductivity as well as to the electric fields which are responsible for the dielectric chain formation and the resulting non-linearity.

It has been mentioned in Section 3(a) that the I–V characteristic passes its inflection point under the slope of the lower tangent:

$$\tan\alpha_0=(i_0/E_0)\tan\beta=\sigma_0\tan\beta. \qquad (4a)$$

The fact that σ_0 depends to a very large degree on the previous history, i.e., on the momentary polarization, makes any pre-calculation very difficult if not impossible.

The deviation of the solid curve from the initial zero tangent can be represented by a cubical parabola, at least at low polarizing fields, in the vicinity of the zero point. Hence, the lower portion of the actual i–E characteristic may be described by:

$$\begin{aligned} i&=\sigma_0E+\sigma_0KE^3 \\ &=\sigma_0E[1+KE^2]. \end{aligned} \qquad (5a)$$

The so-called "field coefficient" K characterizes the non-linearity in first-order approximation.

As soon as the polarizing field becomes increasingly stronger the characteristic remains below the simple parabola. This deviation can be perceived by diminishing K in relation to E:

$$K'=K/(1+PE^2), \qquad (6)$$

where P is an additional quantity which characterizes the non-linearity of the extended characteristic. Hence, Eq. (5a) assumes the more accurate form:

$$i=\sigma_0E[1+KE^2/(1+PE^2)]. \qquad (5b)$$

The evaluation of the measured characteristic in Fig. 10 gives $\sigma_0=2\cdot10^{-7}$ mhos per cm, $K=1.2\cdot10^{-5}$, and $P=9\cdot10^{-7}$.

By means of this analytical representation the condition can be extended to field strengths far above the permissible measuring range. Equation (5b), with the

FIG. 7. (a) Non-linear bridge oscillogram with (b) calibration line.

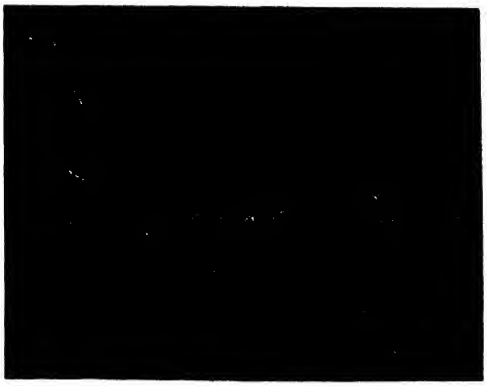


FIG. 8. Non-linear bridge oscillogram at successively increasing driving voltages.

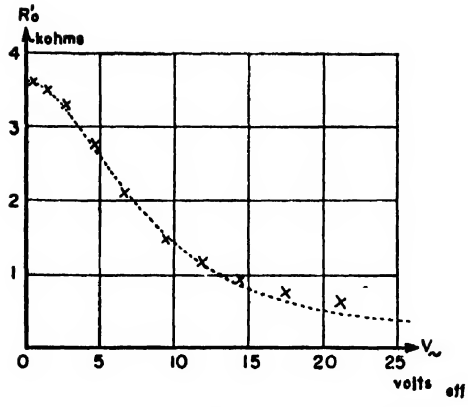


FIG. 9. The zero resistance R_0' in relation to the alternating voltage $V_\sim$ across the colloidal resistor.

experimentally determined values, yields the dashed characteristic shown in Fig. 10. Of special interest is the condition arising when it approaches the upper tangent whose slope follows when $E\rightarrow\infty$ as:

$$\sigma_{sat}=\sigma_0[1+K/P]. \tag{4b}$$

This saturation conductivity is the only quantity which can be understood very simply as characterizing the phenomenon that all semiconductive particles are compressed by infinitely powerful forces of electrostatic attraction in the form of a solid block. In other words, the saturation conductivity (in the present example 13 times the zero conductivity) must be identical with the conductivity of solid graphite, or, for example, coal, multiplied by the mixture ratio of the suspension, because the cube of unity is filled only in part by the graphite, while the rest is filled by the insulating carrier oil.

4. CONDUCTIVITY AND RESISTANCE CHARACTERISTICS

(a) Non-Linear Conductivity

The change in the relative direct current conductivity referred to zero conductivity can be computed easily

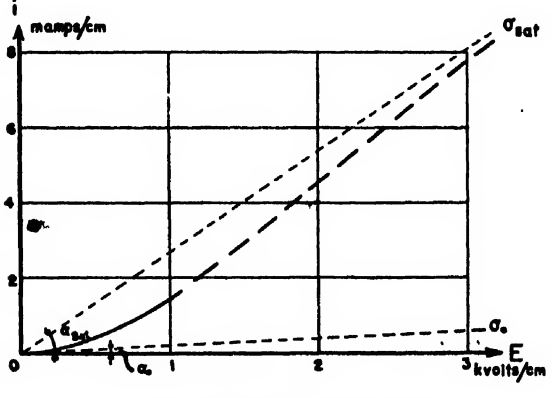


FIG. 10. Analytical I–E characteristic.

from Eq. (5b):

$$\Delta\sigma/\sigma_0=\sigma/\sigma_0-1=K\cdot E^2/(1+P\cdot E^2). \tag{7a}$$

Using the values given above for K and P, this quantity is plotted *versus* E as the solid curve in Fig. 11. Since $P<K$ the curve, in the vicinity of the zero point, is a simple parabola:

$$\Delta\sigma/\sigma_0=K\cdot E^2. \tag{7b}$$

At the field strength:

$$E_i=1/(3P)^{\frac{1}{2}}, \tag{7c}$$

in the present example at 610 v per cm, a point of inflection appears around which the relative direct current conductivity changes linearly with E. At very high field strengths, the relative change in conductivity approaches, asymptotically, a saturation value:

$$(\Delta\sigma/\sigma_0)_{sat}=K/P, \tag{7d}$$

i.e., in the present example, 13.4.

If the conductivity characteristic should be computed from the non-linear bridge oscillogram, it can be done by means of the simple formula:

$$\Delta\sigma/\sigma_0=2\tan\alpha/\tan\beta.$$

Apart from the direct current conductivity, a differential or alternating current conductivity which is proportional to the slope of the I–E characteristic at any arbitrary operating point or at any arbitrary bias field E_B must be included. Differentiation of Eq. (5b) yields the relative differential conductivity, i.e., $\sigma_\sim$ referred to σ_0:

$$\sigma_\sim/\sigma_0=1+KE_B^2\frac{3+PE_B^2}{(1+PE_B^2)^2}. \tag{8a}$$

The dashed curve in Fig. 11 illustrates this relation between the differential conductivity and the polarizing field E_B. In contrast to the solid curve, the latter starts at 1, passes a flat maximum of constant conductivity,

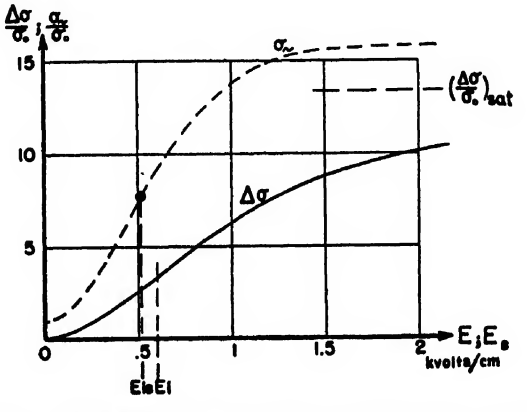


FIG. 11. The relative direct current conductivity $\Delta\sigma/\sigma_0$ and the alternating current conductivity $\sigma_\sim/\sigma_0$ *versus* the polarizing field strength E or E_B.

and approaches asymptotically the same saturation value as before.

At relatively low fields, in the vicinity of the zero point, Eq. (8a) simplifies to:

$$\sigma_\sim/\sigma_0=1+3KE_B^2 \tag{8b}$$

revealing a non-linearity three times greater than in $\Delta\sigma$.

This, at the same time, causes the inflection point to shift to:

$$E_{iB}\sim 1/2(P)^{\frac{1}{2}}, \tag{8c}$$

where the relative differential conductivity

$$(\sigma_\sim/\sigma_0)_i=3KE_\sim/2(P)^{\frac{1}{2}}, \tag{8d}$$

again is proportional to any fluctuations of E superimposed upon the bias field E_{iB}.

(b) Resistance Characteristics

The non-linear conductivity yields the corresponding resistance characteristics. For reasons which will become more apparent later, it is convenient to introduce the differential resistance $R_\sim$ which follows from Eq. (8a) as:

$$R_\sim=\frac{R_0}{1+KE^2[(3+PE^2)/(1+PE^2)^2]}. \tag{9a}$$

If it is considered in relation to the bias voltage V_B, the equation assumes the form:

$$R_\sim=\frac{R_0}{1+KV^2[(3+PV^2)/(e+(P/e)V^2)^2]} \tag{9b}$$

or, if restricted to the permissible range of relatively low voltages:

$$R_\sim=\frac{R_0}{1+3KV^2/(e^2+2PV^2)}. \tag{9c}$$

Examination of this equation reveals that the voltage sensitivity of any colloidal resistor becomes higher the

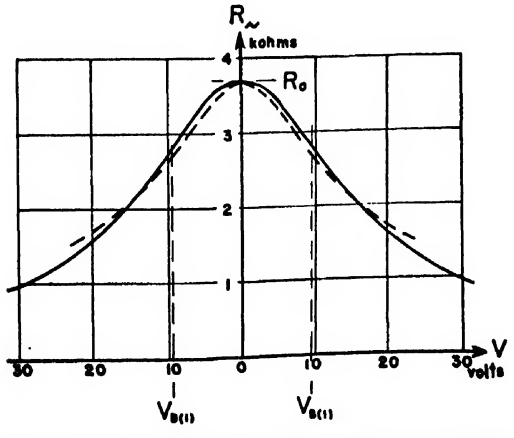


FIG. 12. Non-linear resistance characteristic of a colloidal test resistor with 1-mm electrode separation.

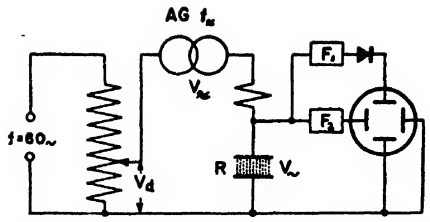


FIG. 13. Arrangement for obtaining oscillograms of differential resistance characteristics with the aid of an auxiliary-frequency $f_\simeq$ superimposed upon the oscillographing voltage $V_\sim$ of 60 cycles.

smaller the oil-gap, or, in first-order approximation, proportional to the second power of e.

As an example, the solid curve in Fig. 12 illustrates the non-linear characteristic of a colloidal test resistor with $R_0=3.7$ kohms and $e=0.1$ cm.

For the following experiments it is convenient to consider only two particular portions of the entire characteristic. First, the maximum portions around R_0 at zero bias which is simply described by:

$$R_{\sim(2)}=\frac{R_0}{1+3(K/e^2)V_\sim^2}. \tag{10a}$$

Second, the linear portions of both side branches around the bias voltage:

$$V_{B(1)}=e/3(K)^{\frac{1}{2}} \tag{10b}$$

in the present example 9.6 volts, at which operating point:

$$R_{\sim(1)}=-1.12R_0eV_\sim(K-P/3)/K^{\frac{1}{2}}$$
$$\sim -1.12R_0V_\sim(K)^{\frac{1}{2}}/e. \tag{10c}$$

The negative sign does not indicate a negative alternating current resistance but signifies only that the resistance decreases if the momentary control voltage $V_\sim$ has the same polarity as the bias voltage.

Naturally, the advantage of a direct oscillography holds also for the resistance characteristics. In order to obtain such characteristics, the arrangement diagrammed in Fig. 13 has been developed to operate as follows.

The basic schematic resembles the previous Fig. 4 with the following modifications: The auxiliary-frequency generator AG superimposes an auxiliary-frequency voltage $V_\simeq$ of some kc upon V_d. The auxiliary-frequency drop across R, after being separated from the 60 cycles by means of the high pass filter F_1 and after rectification between the vertical amplifier output and its deflecting plate, deflects the cathode ray in the Y axis. On the other hand, the horizontal deflection is produced, similarly as before, by the 60-cycle voltage drop $V_\sim$ after the auxiliary-frequency is filtered out by the low pass filter F_2. In this manner the oscilloscope indicates the differential resistance $R_\sim$ in relation to the momentary voltage across R instead of the direct current resistance R. Note that $V_\simeq < V_\sim$ so that only a very short portion of the I–V characteristic is controlled by the auxiliary-frequency.

An example of the differential resistance characteristic of a colloidal test resistor with one mm electrode separation is shown in Fig. 14. Replacing the colloidal resistor with various constant resistors introduces horizontal calibration lines, whereas the sensitivity of the X-deflection was calibrated by means of a known 60-cycle voltage. In this way the resistance oscillogram is completely calibrated. It is redrawn as the dashed curve into Fig. 12, indicating remarkable agreement

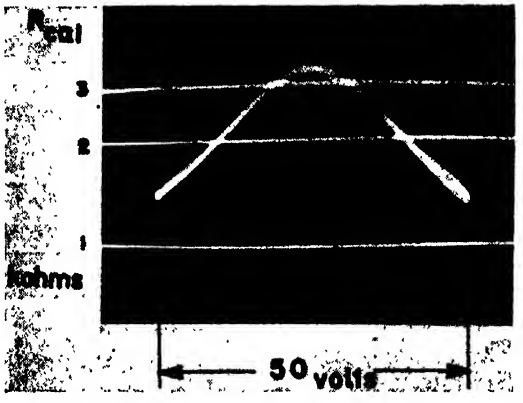


FIG. 14. Oscillographic characteristic of the differential resistance obtained at an auxiliary-frequency of 13 kc, with calibration lines.

with the analytical curve. As a preliminary result of these investigations it can be seen that the colloidal resistor has no drop in sensitivity with the auxiliary-frequency of 13 kc.

5. RECTIFICATION

(a) Pseudo Rectification

The aforementioned turning of the I–V characteristic under the influence of variable alternating voltages finds illustrative expression in the resistance characteristics. This is shown in Fig. 15 where three resistance characteristics are photographed successively at driving voltages $V_\sim$ increasing in increments of three volts. In analogy to Fig. 8 the resistance characteristics, at increasing $V_\sim$, shift downward.

According to these observations R_0 cannot be regarded as constant if the colloidal resistor operates under various voltage amplitudes but depends in turn on the state of polarization and, therefore, on the polarizing alternating field. This peculiar phenomenon can be included in the analysis in two different ways. Assuming that the dashed transfer characteristic in Fig. 16 is controlled by the alternating field $E_\sim = E_0 \cos\omega t$. Then, an average polarization takes place according to the effective field strength, i.e., proportional $E_{eff} = E_0/2^{\frac{1}{2}}$. As a result the initial tangent, whose slope characterizes σ_0, turns into the dotted chord which extends between the effective fields $\pm E_{eff}$ and whose slope indicates a new zero conductivity $\sigma_0' > \sigma_0$.

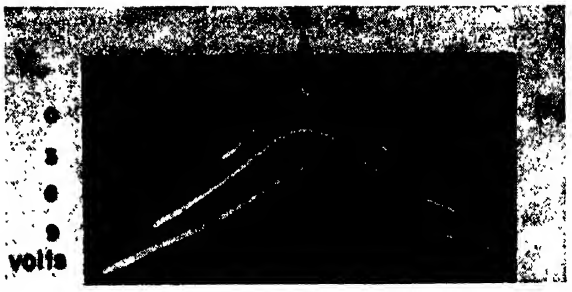


FIG. 15. Non-linear resistance oscillograms at various driving voltages $V_\sim$.

If for the sake of simplicity the simple Eq. (5a) is used, an analytical treatment of this average polarizing effect yields:

$$\sigma_0' = \sigma_0(1+(K/2)E_0^2) = \sigma_0(1+KE_{eff}^2), \qquad (11)$$

where σ_0 denotes the freely adjusting threshold conductivity just above the threshold field.

A more comprehensive explanation of the average polarization is given by considering the non-linear conductivity—or resistance—characteristic being controlled by a large voltage $V_\sim = V_0 \cos\omega t$. This is illustrated in Fig. 17. Here the fluctuating polarization causes a "pseudo-rectification," as shown on the right side of the figure. Because the chain formation is too inert to follow the alternating field itself, R_0 is diminished by the average value ΔR and, therefore, assumes the lower value $R_0' = R_0 - \Delta R$ where R_0 denotes, as before, the threshold resistance just above the threshold voltage. The analytical treatment of this resistive rectification yields, naturally, the same Eq. (11).

An exact example of the relation between R_0' and V_{eff} has already been shown in Fig. 9. There the dotted curve is computed by means of Eq. (11) in the form:

$$R_0' = 3.63/(1+1.51 \cdot V_{eff}^2 \cdot 10^{-2})$$

and is in agreement with the measuring points as well as could be expected.

(b) True Rectification

In addition to the resistive rectification which deals only with an average state of prepolarization, the curved I–V characteristic produces a true rectification. This is explained in greater detail by means of Fig. 18 where the non-linear diode characteristic is controlled by the alternating field $E_0 \cos\omega t$ at various operating points or at various bias fields E_B, respectively. Introducing $E = E_B + E_0 \cos\omega t$ into Eq. (5b) one obtains a rectified current density:

$$i_{rect} = \sigma_0 E_B \frac{1+(K+P)(E_B^2+E_0^2/2)}{1+P(E_B^2+E_0^2/2)}. \qquad (12a)$$

The dashed curves in Fig. 18 represent the true rectifi-

cation at two field amplitudes E_0. Obviously, the diode characteristic becomes erect as in Fig. 8 when subjected to an alternating control field. At very high fields $E_0 \gg E_B$, Eq. (12a) simplifies into:

$$i_{rect} = \sigma_0 E_B (1 + K/P). \qquad (12b)$$

As a result, the rectified diode characteristics straighten and finally approach the dotted line.

In order to prove this result by experiment, the non-linear bridge shown in Fig. 6 is supplemented, as illustrated in Fig. 19. The new arrangement differs from the simple bridge, in that an additional high frequency voltage $V_{\sim}$ of, e.g., 5 mc, is impressed upon the colloidal resistor. This high frequency is supplied by the signal generator "*SG*," whereby *R–F* chokes "*ch*" and capacitors *C'* prevent the high frequency energy from penetrating the bridge. Quite obviously the 60-cycle field $E_{\sim}$ takes the place of the steady bias field E_B which produces, in connection with $E_{\sim}$, the "rectified diode characteristics" shown in Fig. 20.

Whether the true rectification is combined with the previously mentioned pseudo rectification or not, is open to conjecture. Be that as it may, the fact that the diode characteristic indicates no change in frequency up to the 50 mc presently attained, reveals a remarkably good frequency response. The frequency stability marks the colloidal resistor as a useful detector whose versatility will be revealed in greater detail by means of a subsequent paper.

6. NON-LINEAR LISSAJOUS-FIGURES, AND HARMONICS

Up to the present, all oscillograms have been considered only in reference to the characteristic of the colloidal resistor itself. The following section deals with the analysis of the electrical properties of a non-linear

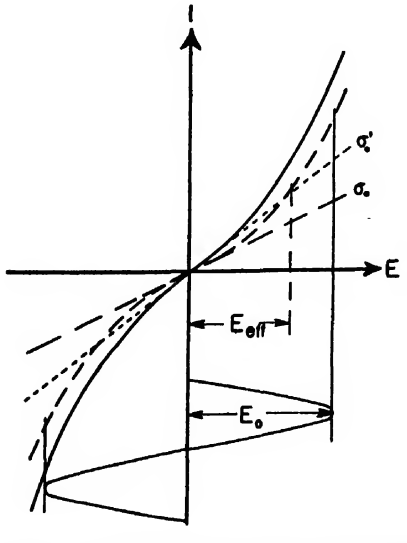


FIG. 16. Average polarization when under the influence of an alternating control field with the amplitude E_0.

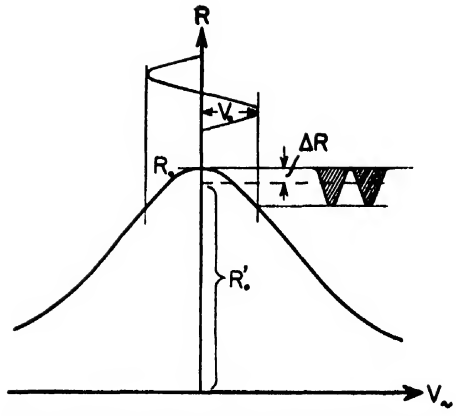


FIG. 17. Illustration of the pseudo rectification.

attenuator when the colloidal resistor operates in series with a constant resistor. For investigation from an electrical point of view, the non-linear bridge already shown in Fig. 6 is obtained, when combined with a linear attenuator.

The foregoing bridge circuit, according to Fig. 21, is modified in that the horizontal input of the oscilloscope is reversed. The *X*-deflection is sinusoidal:

$$D_x = X V_d/2,$$

whereas the diagonal voltage across the middle branch of the bridge is:

$$V_{dg} = V_{\sim} - V_d/2$$
$$= (V_d/2)\{[2/(1+RG)]-1\}.$$

Furthermore, an adjustable bias voltage V_B is added, insuring various operating points of the colloidal resistor.

For the sake of simplicity, the following analysis is restricted to two specific cases: First, the colloidal resistor operates under first-order non-linearity on linear branches of its conductance characteristic at a certain

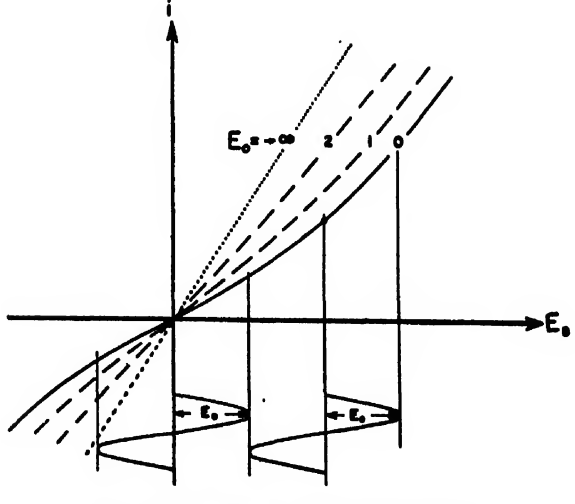


FIG. 18. "Rectified" diode characteristics.

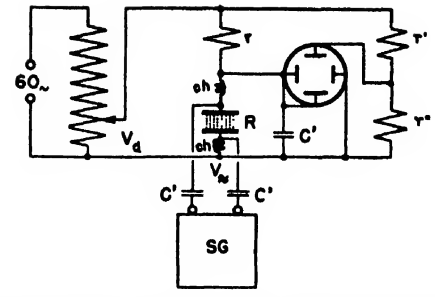


FIG. 19. Schematic of the non-linear bridge with superimposed high frequency voltages for oscillographing rectified diode characteristics.

bias voltage V_{Bi}; second, it operates under the non-linearity of second order, i.e., on the parabolic middle branch without bias.

In case one, the conductance characteristic is linear, following Eq. (7c) at the bias:

$$V_{Bi}=e/(3P)^{\frac{1}{2}}$$

with the slope:

$$G_B/V_d=\pm 3G_B eK/2(P)^{\frac{1}{2}}$$

so that the actual conductance at V_B becomes:

$$G_{\sim}=G_B[1\pm 3eKV_{\sim}/2(P)^{\frac{1}{2}}]$$

where G_B is the direct current conductance at the bias V_{Bi}. Introducing this expression into Eq. (13) and, if the bridge is balanced by means of $rG_B=1$, then:

$$V_{dg}=\frac{V_d}{2}\left[\frac{1}{1\pm eKV_{\sim}/3(P)^{\frac{1}{2}}}-1\right]. \qquad (14a)$$

A simple solution is possible only at very low driving voltages. Then the distortions of the voltage $V_{\sim}$ across R compared to the fundamental amplitude can be neglected. More accurately it means that only distortions of the first order should be considered as distortions caused by the fundamental voltage $V_{\sim}$ alone, and not distortions of the higher order caused by a cross-modulation. According to this simplification $V_{\sim}=V_d/2$. With this value, Eq. (14a) in first-order approximation becomes:

$$V_{dg}=eKV_d/12(P)^{\frac{1}{2}}$$

or furthermore, if V_d is introduced in the form $V_0 \sin\omega t$:

$$V_{dg}=eKV_0^2/24(P)^{\frac{1}{2}}\pm eKV_0^2/24(P)^{\frac{1}{2}}\cos 2\omega t$$
$$=\Delta V_{dg}\pm V_{dg\sim}. \qquad (14b)$$

Apart from a direct voltage ΔV_{dg}, corresponding to the aforementioned rectification, a second harmonic appears whose phase reverses with the polarity of the bias.

Plotting V_{dg} *versus* V_d produces a non-linear Lissajous-figure as shown in Fig. 22(a). If the polarity of the bias is reversed, the oscillogram inverts and assumes the shape (b) illustrating and thus proving the reversal of phase.

The Lissajous-figures are pure parabolas only if taken at low driving voltages. In the present examples, however, the driving voltage was relatively high, so that the parabolas are distorted, partly because the linear portions of the conductivity characteristic are overcontrolled, and partly because of the cross-modulation. The linear-time oscillogram shown in Fig. 22(c) reveals the additional distortion of the original second harmonic and indicates a full series of higher order harmonics predominantly of even order.

The second example of second-order non-linearity at zero bias yields, after introducing:

$$G_{\sim}=G_0(1+K(V_{\sim}/e)^2)$$

into Eq. (13), when limited again to low voltages, with $V_d=V_0 \sin\omega t$, and $V_{\sim}=V_d/2$:

$$V_{dg}=-3KV_0^3[\sin\omega t-(\sin 3\omega t)/3]/64e^2. \qquad (15)$$

Obviously the second-order non-linearity, characterized by the second power of the control voltage, causes a fundamental voltage to appear—because of true rectification—together with a harmonic of the second-plus-one, i.e., the third-order harmonic. Plotting V_{dg} *versus* V_d produces a Lissajous-figure similar to the non-linear bridge characteristic shown in Fig. 7. The fundamental can be suppressed by rebalancing the bridge by means of a slightly diminished r. At this rebalanced condition, only the third harmonic remains and the Lissajous-

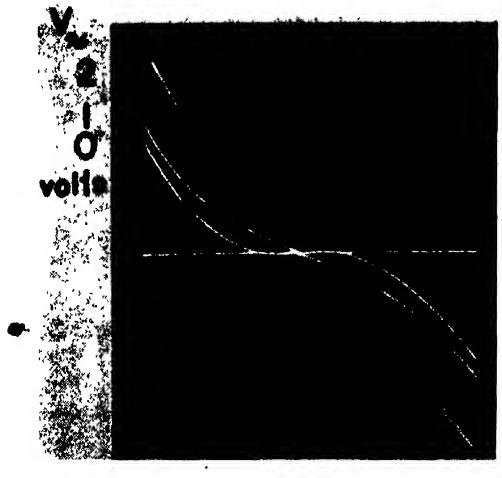


FIG. 20. Rectified diode characteristics obtained with an oscillographing frequency of 60 cycles superimposed by 5 mc.

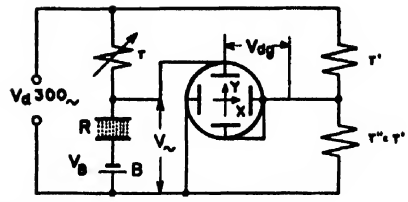
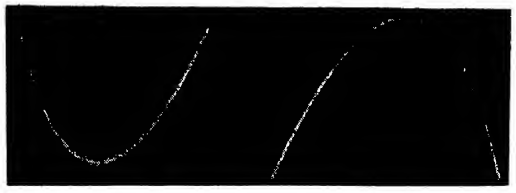


FIG. 21. Schematic of arrangement for producing the non-linear Lissajous figures shown in Figs. 22 and 23.



(a) (b)

(c)

FIG. 22. Non-linear Lissajous figures at first-order non-linearity.

figure lies symmetrically on the X axis as shown in Fig. 23(a).

Here again, the oscillogram has been taken at relatively high driving voltages so that the validity of the developed formulas is exceeded. Consequently, the cross-modulation produces higher harmonics of an odd order as can be seen in the linear time oscillogram (b).

CONCLUSIONS

The concept of the formation of purely dielectric or of additionally semiconductive chains and the resulting non-linearities of dielectric constant or conductivity, respectively, may be important for understanding the philosophy of Seignette-electricity and some peculiarities of semiconductors.

On the other hand, the operational characteristics of the colloidal resistors assure certain interesting applications, as, for instance, new types of detectors, frequency converters and mixers, stabilizers and limiters for d.c. and a.c. voltages, linearizers for vacuum tube amplifiers, etc., where their large range of voltage sensitivity and power dissipation reveals entirely new possibilities. Furthermore, the sensitivity of the dielectric-semiconductive chains to mechanical or quasi-mechanical influences makes it possible to design sensitive conversion elements and transducers. In addition, when the dielectric-semiconductive particles are also ferromagnetic, magnetic forces of attraction can be superimposed upon the electrostatic attractions so that the resistance

(a)

(b)

FIG. 23. Non-linear Lissajous figures at zero bias.

of such a ferromagnetic resistor can be controlled in a manner similar to Hull's first magnetron. It remains for further experiments to prove whether or not it is possible to substitute the magnetic control field by an electric field and to develop liquid-filled amplifier and oscillator devices similar to the transistor.

During the preparation of this paper, it was possible to change from the liquid-filled colloidal resistors to dry types (the so-called polar resistors or "*Polaristors*") displaying basically all the properties mentioned before. All of these problems are under study and results of the investigations will be published in the near future.

ACKNOWLEDGMENTS

The author is indebted to the Acheson Colloids Corporation for furnishing him samples of Oildag, Castordag, and Glydag. Furthermore, the author wishes to extend his sincere thanks to Dr. Ernst O. H. Friederich for his efficient and able assistance in producing various samples of dielectric, semiconductive, and ferromagnetic materials. Last but not least the writer is very grateful to Mr. Harold Berman and Mr. Fred A. Koether for their valuable help in preparing this paper.

Internal Friction of Metals at Very High Temperatures*

T'ING-SUI KÊ
Institute for the Study of Metals, University of Chicago, Chicago, Illinois
(Received October 10, 1949)

In connection with the study of the internal friction peak (*versus* temperature) associated with the viscous behavior of grain boundaries in metals, some other effects were observed at higher temperatures causing an additional internal friction in superposition with the high temperature branch of this internal friction peak. This additional internal friction at high temperatures was found to have its origins in some effects of cold-working introduced in the interior of the grains which remains even after the complete recrystallization of the cold-worked specimens. It increases with the amount of cold-work the specimen was subjected to before and after recrystallization; it decreases with annealing at successively higher temperatures until a stable value is reached; and it increases with the precipitated impurity content in the specimen. This high temperature internal friction was found to be very high in an aged specimen of high purity aluminum alloyed with 4 percent of copper. These observations are consistent with the viewpoint that this internal friction is caused by the presence of dislocations in the interior of the specimen although the mechanism of giving rise to the internal friction is unknown. It is pointed out that a study of high temperature creep under very low stress may be conveniently carried out through the internal friction measurements described.

INTRODUCTION

IN the study of the high temperature internal friction of annealed polycrystalline aluminum, a peak has been observed when the internal friction is plotted as a function of temperature of measurement.[1] Such a peak was not observed in single crystal aluminum under similar conditions. Extensive study has shown that the internal friction peak is associated with the viscous behavior of the grain boundaries in metals. However, it has also been observed that the high temperature branch of this internal friction peak did not drop as rapidly as expected if this peak is entirely due to the viscous behavior of grain boundaries (Fig. 2 of reference 1). Apparently, some other effects occur at higher temperatures causing an additional internal friction in superposition with the high temperature branch of the grain boundary internal friction peak. The physical origin of this additional internal friction at very high temperatures is not understood.

The purpose of this paper is to report the results of a study on the factors influencing this additional internal friction. The physical origin of this internal friction will be discussed. All the internal friction measurements were made using a torsion pendulum as previously described.[1] The frequency of vibration was about one cycle per second. The applied stress was sufficiently small so that the magnitude of internal friction is independent of stress level at each temperature range concerned.

The additional internal friction at high temperatures has been observed in a number of metals in addition to aluminum. Figure 1 shows the internal friction peaks associated with the viscous behavior of grain boundaries for Mg,[1] Al,[1] Cu,[2] alpha-brass,[3] and Fe.[4] They were all measured in torsional vibration with a frequency of about one cycle per second. In each case it is seen that the internal friction increases again at temperatures higher than the optimum temperature of the grain boundary internal friction peak. Such a feature has also been observed by Köster in transverse vibration for Cu and Ag.[5] The frequency of vibration he used was 795 cycles per second for Cu and 642 cycles per second for Ag. It seems thus the occurrence of additional internal friction at high temperatures is independent of the mode of vibration and is a common feature of all metals.

This high temperature internal friction cannot be caused by the strain amplitude effect of the applied stress since, as we have mentioned above, the applied stress has been always kept small enough for each temperature so that the internal friction is independent of stress level. In making internal friction measurements with a torsion pendulum, a small load including the

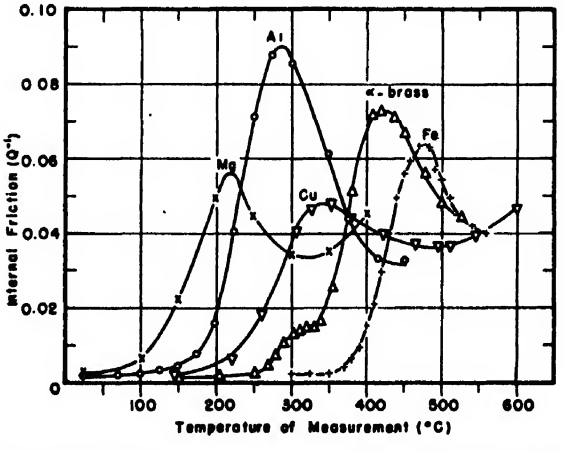


FIG. 1. Internal friction peaks (*versus* temperature) associated with the viscous behavior of grain boundaries in several metals.

* This research has been supported in part by ONR (contract No. N-6ori-20-IV, NR 019 302).
[1] T. S. Kê, Phys. Rev. 71, 533 (1947).
[2] T. S. Kê, J. App. Phys. 20, 274 (1949).
[3] T. S. Kê, J. App. Phys. 19, 285 (1948).
[4] T. S. Kê, "Metals technology," Tech. Pub. 2370 (June, 1948).
[5] W. Köster, Z. Metallkde. 39, 9 (1948).

inertia arm was applied longitudinally to the specimen. The effect of this longitudinal load has been tested for each metal. Figure 2 shows the internal friction for the case of iron when two different longitudinal loads were used in the measurement. It is seen that the whole curve is similar in both cases although one longitudinal load was seven times heavier than the other. Consequently, the additional internal friction at elevated temperatures is not attributable to the longitudinal load.

INFLUENCE OF COLD-WORK

We have mentioned previously that the high temperature internal friction is quite sensitive to the amount of cold-work the specimen was subjected to before recrystallization. Figure 3 shows the internal friction curves of specimens of recrystallized 99.991 aluminum: Curve I for the case when the specimen was subjected to a cold-work of 70 percent reduction in area and Curve II a cold-work of 95 percent reduction in area. Both specimens were recrystallized after being annealed at 450°C for five hours and the grain sizes of both were found to be similar. It is seen that the high temperature branch of the curve is lower in the case of the specimen subjected to a smaller amount of cold-work prior to recrystallization. This indicates that a smaller amount of prior cold-work causes a smaller addititional internal friction at high temperatures.† Curve III shows the internal friction curve of a specimen which was subjected to a small amount of cold-work after recrystallization. This cold-work has modified considerably the high temperature branch of the internal friction peak. It is apparent, thus, that cold-working both prior to and after recrystallization has the effect of raising the additional internal friction at elevated temperatures.

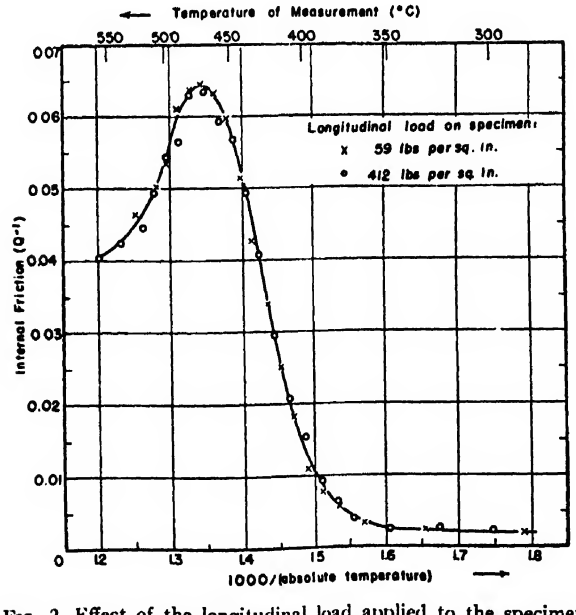


FIG. 2. Effect of the longitudinal load applied to the specimen upon the high temperature internal friction.

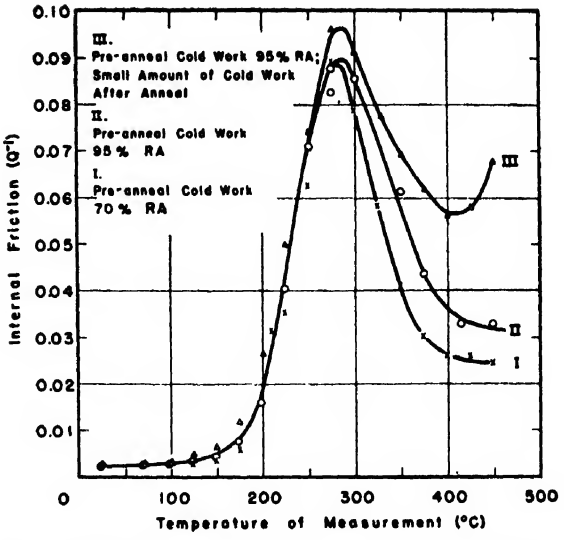


FIG. 3. Influence of the cold-work performed on the specimen before and after recrystallization.

This additional internal friction has been found to decrease as the annealing temperature goes successively higher. Figure 4 shows the effect of annealing on 99.991 aluminum subjected to a prior cold-work of 70 percent reduction in area. Figure 5 shows the case of aluminum of commercial purity (Al 2S) subjected to a prior cold-work of 96 percent reduction in area. It is to be pointed out that such a decrease of internal friction with an increase of annealing temperature reaches eventually a steady state which tendency can be seen by comparing the 400°C and 500°C annealing curves in Fig. 5.

ORIGIN OF THE HIGH TEMPERATURE INTERNAL FRICTION

The high temperature internal friction described above has also been observed in single crystal aluminum (99.6 percent purity) as shown by Curve IV in Fig. 6. The internal friction peak associated with the viscous behavior of the grain boundaries is absent in this case. However, the internal friction is seen to rise at very high temperatures. Curves I–III of Fig. 6 are the internal friction curves of several similar aluminum single crystal wires which were subjected, respectively, to the amount of cold-work shown and were then annealed at 450°C for one hour. The high temperature internal friction is seen to increase with the amount of cold-working. This indicates that this internal friction had originated from some effects of cold-working introduced inside the grains.

† In the present case the position of the grain boundary internal friction peak did not shift with the different amount of prior cold-work (70 percent and 90 percent reduction). The situation is different when the amount of prior cold-work involved is smaller (see Fig. 2 of reference 2).

INFLUENCE OF IMPURITIES

A careful study of the internal friction curves shown in Fig. 1 revealed one interesting feature: The rise of internal friction at elevated temperatures is relatively higher when the grain boundary internal friction peak is relatively lower. Thus, in the case of aluminum, the grain boundary peak is the highest and the relative rise of the high temperature internal friction is the lowest. Previous experiments have shown that the viscous slip along grain boundaries can be partly or completely blocked by the presence of impurities at the grain boundaries. A lower grain boundary internal friction peak thus indicates a higher impurity content

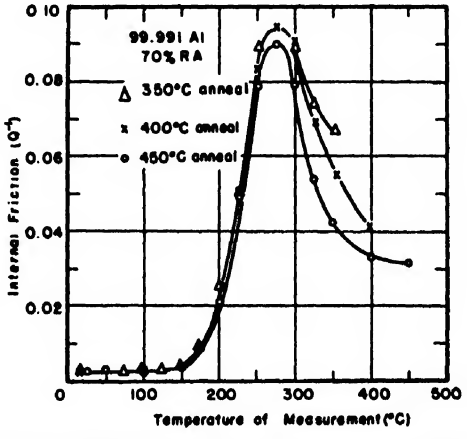


FIG. 4. Influence of annealing at successively higher temperatures (99.991 Al).

at the grain boundaries. The observation that the high temperature internal friction is larger whenever the grain boundary internal friction is smaller suggests that the former is connected with, or at least is influenced by, the presence of impurities.

Figure 7 shows the internal friction curves of polycrystalline aluminum of varying impurity content. Curve I is for high purity 99.991 percent aluminum; Curve II for aluminum 2S (containing as impurities 0.14 percent Cu, 0.48 percent Fe, 0.11 percent Si); Curve III for high purity aluminum alloyed with four percent of high purity copper. The prior cold-work and the subsequent annealing were similar in all cases. The grain sizes of the three specimens are also similar. The differences in these curves shown in Fig. 7 are thus attributable to the different amounts of impurity content. It is seen that in the case of aluminum containing 4 percent of copper, the grain boundary peak has been completely eliminated while the internal friction at 450°C is almost three times larger than the case when no copper was added.

STUDY OF AL-CU ALLOYS

In order to find out how the impurity content influences the internal friction at elevated temperatures, a detailed study has been made for the case of high

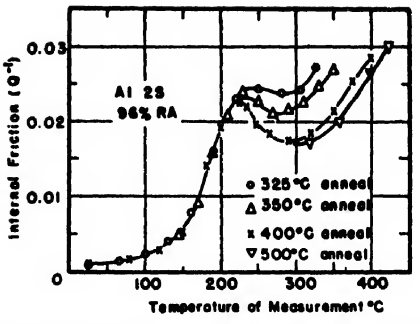


FIG. 5. Influence of annealing at successively higher temperatures (aluminum 2S).

purity aluminum alloyed with four percent of copper. In this case the impurity responsible for the effect concerned is undoubtedly copper. Figure 8 illustrates how the internal friction changes with heat treatment. The specimen was first solution-treated at 525°C for one hour and was then cooled slowly from 525°C so that the copper precipitates out in some form from the solid solution. It is seen that, after this treatment, the internal friction peak associated with the viscous behavior of the grain boundaries was eliminated while the additional internal friction at elevated temperatures is unusually high. The elimination of the grain boundary peak is supposedly due to the precipitation of copper in some form at the grain boundaries. As we have shown in the last section that the high temperature internal friction is not attributable to the grain boundaries, the rise of this internal friction is probably connected with the general precipitation of copper in aluminum, that is, the precipitation at the interior of the grains. The curve marked with (1) shows the case when the specimen was quenched from 525°C into water at room temperature. Under this treatment the

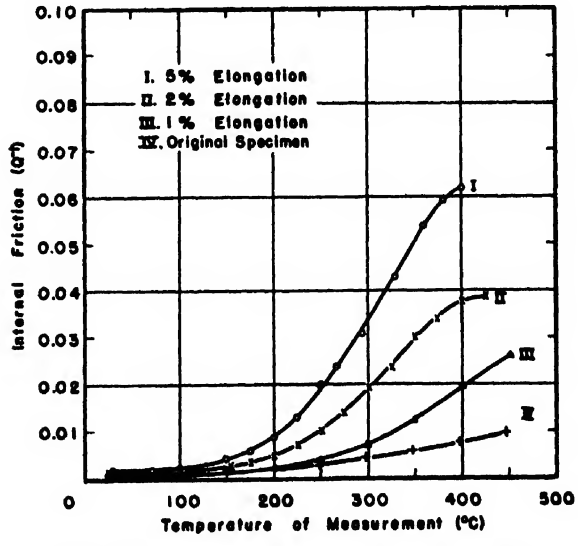


FIG. 6. High temperature internal friction of single crystal aluminum.

copper remains in solid solution in a supersaturated form. However, as the specimen is tempered at a higher temperature, the precipitation of copper will proceed at an increasing rate. The portion of curve marked (1) gives the grain boundary internal friction peak around 200°C which was measured by heating the specimen rapidly from room temperature to 200°C. This peak is fairly high although it is much lower than that for high purity aluminum without copper. This indicates that during the heating period some precipitation has taken place at the grain boundaries to block partly the viscous intercrystalline slip. When the specimen was heated rapidly from 200°C to 450°C, the internal friction at

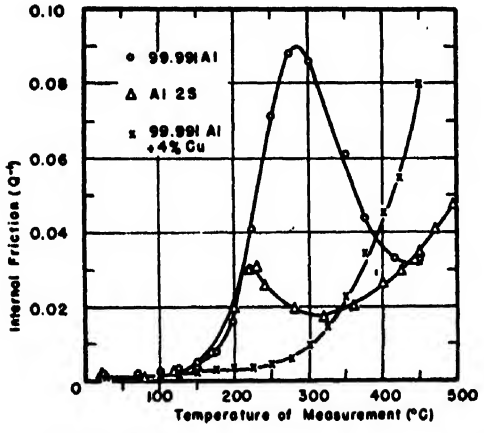


FIG. 7. Influence of impurity content upon the high temperature internal friction of aluminum.

450°C was measured and is given by (1). This internal friction was found to increase with time and its magnitude is shown by the time sequence of (1), (2), (3). This increase seems to be due to the fact that more copper was precipitated out from solid solution by tempering at 450°C. After the time sequence (3), the specimen was rapidly cooled from 450°C to 200°C. The internal friction at 200°C as given by (4) is much smaller than the original curve marked (1), indicating that more precipitation had been taking place at the grain boundaries during the tempering at 450°C. As the grain boundary precipitation is accompanied by general precipitation especially when the specimen is tempered at elevated temperatures, it is natural to assume that the rise of internal friction at 450°C by tempering is associated with the general precipitation of copper.

When the specimen was cooled slowly from 525°C, the solution treatment temperature, to 450°C, the precipitation of copper has probably been completed and the amount of copper left in solid solution is determined by the solubility of copper in aluminum at 450°C which is 2.60 percent by weight.[6] Under this condition the

[6] E. H. Dix, Jr. and H. H. Richardson, Trans. A.I.M.E. 73, 560 (1926); Brown, Fink, and Hunter, Trans. A.I.M.E. 143, 115 (1941).

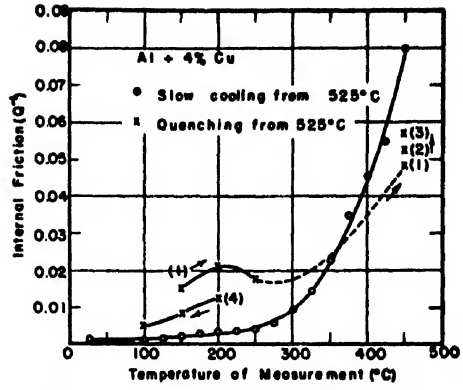


FIG. 8. Influence of heat treatment and the state of precipitation upon the high temperature internal friction of Al containing 4 percent of Cu.

internal friction at 450°C was found to be the highest as shown in Fig. 8. This indicates that the internal friction is the highest when the copper is in the precipitated state of some form.

The internal friction at elevated temperatures has been found to be independent of the impurities in dissolved form. Figure 9 shows the internal friction curves of high purity aluminum and of high purity aluminum alloyed with ½ percent of high purity copper. At 300°C the solubility of copper in aluminum is 0.45 percent.[6] The ½ percent copper in the specimen is thus in dissolved form at temperatures a little over 300°C. It is seen that the portions of the two curves over 300°C are almost identical, indicating that the ½ percent Cu in dissolved form has nothing to do with the internal friction at elevated temperatures. At temperatures lower than 300°C, the internal friction curve for the specimen containing ½ percent of copper is lower. This is evidently because of the blocking of grain boundary slip by copper precipitated in some form at the grain boundaries at temperatures lower than 300°C.

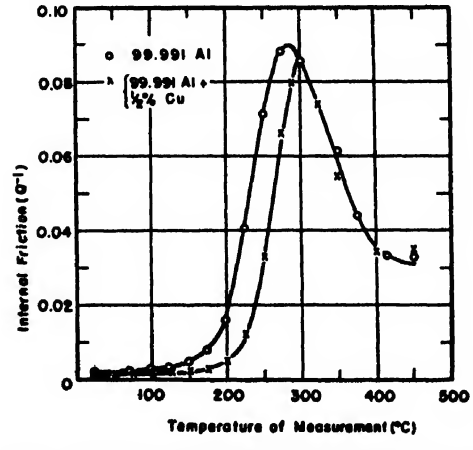


FIG. 9. Influence of dissolved impurity upon the high temperature internal friction.

TABLE I. Chemical and spectrographic analyses of aluminum specimens.

Alloying elements	99.991 Al	Aluminum 2S	Chemically pure Al
Cu	0.002%	0.14%	0.13%
Fe	0.002	0.48	0.053
Si	0.002	0.11	0.18
Mg	0.003		

STUDY OF TERNARY ALUMINUM ALLOYS

In describing the effect of impurity upon the internal friction at elevated temperatures, we have mentioned in connection with Fig. 7 that the high temperature internal friction of aluminum increases with the amount of copper contained in the specimen. Thus aluminum 2S containing 0.14 percent of copper has a higher internal friction than that of high purity aluminum but a lower internal friction than that of high purity aluminum alloyed with 4 percent of copper. A closer examination indicates that in the case of aluminum 2S, the situation is not so simple as the solubility of copper in aluminum at 450°C is 2.60 percent. The 0.14 percent of copper in the specimen should be in solid solution at 450°C and the high temperature internal friction should be small at this temperature. Figure 9 shows this is the case for high purity aluminum alloyed with $\frac{1}{2}$ percent of copper. The anomalous condition in aluminum 2S indicates that a ternary or more complicated system is involved. The situation is more puzzling in the case of chemically pure aluminum containing 0.13 percent of copper, which gives an internal friction at 450°C as high as 0.11 (for a prior cold-work of 77 percent reduction in area. See Fig. 3 of reference 2). This internal friction is even higher than that of aluminum alloyed with 4 percent of copper at the same temperature. Table I gives the impurity content of the high purity aluminum and other samples.

It is seen that the copper content in aluminum 2S and in the chemically pure Al is about the same. The other impurities are chiefly iron and silicon. An insoluble Al-Cu-Fe phase or Al-Fe-Si phase may be formed and precipitates out from the solid solution although the copper and silicon, if present alone, will go into solid solution at the temperature range concerned. If this is the case, then the rise of the high temperature internal friction may be attributed to the presence of these ternary phases.

However, this still cannot explain why the internal friction of the chemically pure Al at 450°C is much higher than that of aluminum 2S. Figure 10 shows that at 450°C the internal friction of the former is higher than that of the latter although the previous cold-work in the former is only 42 percent reduction in area while it is 96 percent in the latter. It is not understood whether this difference can be attributed to the difference of the iron content in the two specimens (see Table I) or to some other unknown factors such as the past history of the specimens.

DISCUSSION OF RESULTS

In summary, we have found that the additional internal friction which dominates at elevated temperatures has the following features:

1. It occurs in polycrystalline and single crystal metals.
2. It increases with the amount of cold-work the specimen was subjected to before and after recrystallization.
3. It decreases with annealing at successively higher temperatures until a stable value is reached.
4. It increases with the precipitated impurity content in the specimen.

Let us examine now what are the underlying mechanisms of this high temperature internal friction.

A glance at Fig. 1 shows that the internal friction at elevated temperatures always occurs after the grain boundary relaxation has been completed. As the shear stress is relaxed along the grain boundaries, a concentration of stress will be built up at the grain corners.[7] Under proper conditions the local stress concentration may be large enough to initiate slip at the grain corners.[8] If this does happen it will certainly cause a considerable increase of internal friction. As a matter of fact, Hanson and Wheeler[9] have observed the macroscopic movement of grain boundaries at elevated temperatures. However, this was the case only when the applied stress is large. In our present case the applied stress is small, so the stress concentration at the grain corners is probably not large enough to initiate slip. This seems to be true as the curves in Fig. 7 show that the internal friction at 450°C is high in spite of the fact that the grain boundary relaxation was blocked by the presence of impurities.

If slip at grain corners has been initiated during measurements at 450°C, the internal friction should be sensitive to the stress amplitude. On the contrary, the internal friction has been found to be independent of stress levels under the experimental conditions.

In his study of amorphous plasticity, Orowan has pointed out that a disordered arrangement of atoms leads to a viscous type of flow if the temperature is

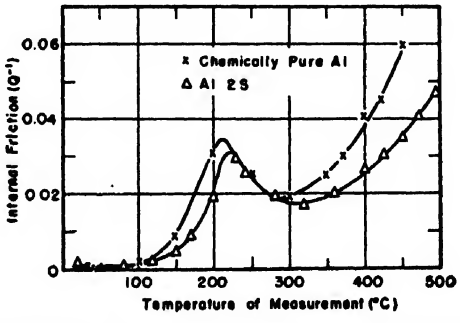


FIG. 10. High temperature internal friction of aluminum 2S and of chemically pure aluminum.

[7] C. Zener, *Elasticity and Anelasticity of Metals* (The University of Chicago Press, 1948), p. 127.
[8] C. Zener, Phys. Rev. 69, 128 (1946).
[9] D. Hanson and M. A. Wheeler, J. Inst. Metals 45, 229 (1931).

sufficiently high.[10] "Atomic disorder can be found not only at grain boundaries, but also inside the crystals if these have been plastically distorted, and particularly during recrystallization and phase transformations when the atoms leave one lattice and join another."[11] Along this line of thought this author has recently described a grain boundary model in which the transition region at the boundary was considered as consisting of numerous disordered groups of atoms or diffused holes.[2] The viscous intercrystalline slip was considered to occur through the atomic rearrangement by thermal agitation within each "disordered group" by a shear process involving as units of flow only a few atoms. Such a grain boundary model and slip mechanism are consistent with experimental facts. Among others, such a viscous intercrystalline slip will give rise to internal friction.

We have also pointed out that the "disordered groups" of atoms may conceivably exist also in the interior of the grains, except that they are much smaller in number than those existing at the grain boundaries. However, it can be shown that such disordered groups of atoms, when existing isolated in the interior of grains, cannot cause appreciable stress relaxation to account for the high internal friction observed at high temperatures as shown in Fig. 7 in which the internal friction of Al+4 percent of Cu is shown to be as high as 0.08 at 450°C. In the language of dislocation theory, we can say that numerous dislocations were introduced in the specimen during the cold-working and some dislocations remain therein even after the complete recrystallization of the specimen. The stress-induced movement of the dislocations can give rise to internal friction although the exact mechanism involved is not known.

The number of dislocations introduced in the specimen increases with the amount of cold-work. It is conceivable that the number of dislocations remaining in the specimen after recrystallization is larger in the case of a larger prior cold-work. The number of survival can be further reduced by annealing at a higher temperature until an equilibrium state is reached. In each case the internal friction increases with the number of dislocations existing in the specimen.

The effect of undissolved impurities on the observed internal friction may be primarily due to their retarding effect upon the removal of dislocations by annealing. This explains the observed facts that the high temperature internal friction is raised by undissolved impurities. However, additional consideration must be given to the complicated influences of impurities upon the stress-induced movement of dislocations.

[10] Ing. E. Orowan, Proc. West of Scotland Iron and Steel Inst. (February, 1947).

[11] R. Becker, Zeits. f. techn. Phys. 7, 547 (1926).

CONCLUDING REMARKS AND CORRELATION WITH CREEP MEASUREMENTS

The foregoing discussions indicate that the experimental observations are consistent with the viewpoint that the internal friction at elevated temperatures under small stresses (of non-grain boundary origin) is caused by the presence of dislocations in the interior of the specimen. Accordingly, any factor that influences the number and the movement of dislocations in the specimen will result in a change of this internal friction. Such factors consist of the past thermal and mechanical treatments on the specimen, the impurity content, and the state of the impurities in the specimen.

Since a rise of internal friction has been shown to correspond to a rise of the rate of anelastic creep under low stress,[1,7] the additional internal friction at elevated temperatures may be correlated with the anelastic creep of metals at elevated temperatures in the same manner as the grain boundary internal friction was correlated with the anelastic creep of metals originating from the viscous behavior of grain boundaries (grain boundary creep).[1] Consequently, the discussions concerning the physical aspects and origin of the high temperature internal friction hold equally well for the case of anelastic creep under very small stress.

In conventional creep experiments the load applied to the specimen is usually large. This complicates the situation since the applied stress introduces additional creep properties to the specimen under test. To understand the physical aspects of the creep of the original specimen, the creep experiments must be done under sufficiently small stress so that the applied stress furnishes only a directive effect. A step toward this direction has been made by Chalmers in his study of the effect of applied stress upon the micro-creep of tin.[12] The study of creep through measurements of internal friction at low stress, as described in this paper, may prove to be both expedient and capable of giving unambiguous results for theoretical analysis.

[12] B. Chalmers, Proc. Roy. Soc. (A) **156**, 427 (1936); J. Inst. Metals **61**, 103 (1937).

The Orientation Dependence of the Rate of Grain Boundary Migration*

PAUL A. BECK, PHILIP R. SPERRY, AND HSUN HU
Department of Metallurgy, University of Notre Dame, Notre Dame, Indiana
(Received November 14, 1949)

Very small recrystallized grains of a large variety of orientations were produced locally in lightly rolled high purity aluminum crystals. The rate of growth of these small grains varied greatly, depending on their relative orientation with respect to the matrix crystal. Grains with an orientation corresponding to a [111] rotation of about 40° showed by far the highest rate of growth. These results indicate that recrystallization textures developing in the spontaneous recrystallization of deformed single crystals can be accounted for by the orientation dependence of the rate of growth.

THE large grains growing in an essentially strain-free fine-grained matrix of single orientation texture (coarsening) have quite definite orientation relationships with the matrix. In face-centered cubic metals, such as aluminum,[1] copper,[2,3] and nickel-iron,[4] the orientation relationship most frequently found corresponds to a rotation of about 30–45° around a [111] axis.

It was recently observed[1] that for the essentially strain-free grains growing in a highly oriented cold-worked matrix (rolled or compressed single crystal of aluminum), where the driving force of grain boundary migration is the strain energy stored in the cold-worked metal (recrystallization), the orientation relationship with the matrix is the same as for coarsening.[5] The identity of the orientation relationships in the two processes is all the more remarkable as the driving energy for recrystallization after large deformations is estimated to about 100 or 1000 times that for coarsening.

FIG. 1. Strain induced boundary migration in high purity aluminum. Area swept by boundary while moving from 1 to 2, during last anneal, is essentially strain-free, has same orientation as cold-worked grain *A*. Oblique illumination. Magnification 75×.

The understanding of these reorientation phenomena depends largely on the question whether the orientation relationship is determined by the availability of nuclei in definite orientations only, or whether it is determined by the dependence of the rate of growth on the orientation. Oriented nucleation was proposed by Burgers and Louwerse[6] for recrystallization and by Burgers[7] and recently by Kronberg and Wilson[3] for coarsening. Oriented growth in coarsening was suggested by van Arkel.[8] The possibility of interpreting the reorientation in recrystallization on the basis of an oriented growth theory, was pointed out by Barrett.[9] Recent work by three independent groups of investigators,[1,3,4] with three different metals, furnished overwhelming evidence to the effect that in coarsening "nucleated" by small grains, formed amidst the highly oriented matrix in a large variety of orientations by recrystallization after inhomogeneous local deformation (sawing or shearing of the edges, ball indentation, needle scratch), the grains growing to large size have the same orientation relationship to the matrix as the coarse grains formed spontaneously (without artificial nucleation). These experiments prove that the rate of growth in coarsening does depend on the orientation and that the orientations favored in spontaneous coarsening are actually ac-

* This work was supported by the ONR, U. S. Navy, Contract No. N6ori-165, T.O. 1.

[1] P. A. Beck and Hsun Hu, "Recrystallization texture and coarsening texture in high purity aluminum," Trans. A.I.M.E. 185, 627 (1949).

[2] J. S. Bowles and W. Boas, "The effect of crystal arrangement on secondary recrystallization in metals," J. Inst. Metals 74, 501 (1948).

[3] M. L. Kronberg and F. H. Wilson, "Secondary recrystallization in copper," Trans. A.I.M.E. 185, 501 (1949).

[4] G. W. Rathenau and J. F. H. Custers, "Secondary recrystallization of face-centered Ni-Fe alloys," Philips Research Reports 4, 241 (1949).

[5] Unpublished work, carried out at this laboratory, gave similar results for the recrystallization of rolled copper single crystals. Maddin, Mathewson, and Hibbard ["The origin of annealing twins in brass," Trans. A.I.M.E. 185, 655 (1949)] found similar orientation relationship for the recrystallization of extended brass single crystals, and Kronberg and Wilson (see reference 3) for the recrystallization of lightly rolled polycrystalline copper with cube texture.

[6] W. G. Burgers and P. C. Louwerse, "Über den zusammenhang zwischen deformationsvorgang und rekristallisationstexture bei aluminium," Zeits. f. Physik 67, 605 (1931).

[7] W. G. Burgers, "Rekristallisation, Verformter Zustand und Erholung" (1941).

[8] A. E. van Arkel, "Quelques phenomenes de recristallisation," Rev. d. Met. 33, 197 (1936).

[9] C. S. Barrett, "Recrystallization texture in aluminum after compression," Trans. A.I.M.E. 137, 128 (1940).

counted for by their high rate of growth.† It is significant that no observation was so far reported by any investigator that could not be readily based on oriented growth. Finally, Dunn reported[10] the variation of the rate of growth with orientation for coarsening in silicon ferrite. There has been, however, no similarly direct evidence available to prove that in recrystallization, too, the rate of growth does in fact depend on the orientation of the growing grain with regard to the matrix.

The chief difficulty in carrying out a recrystallization experiment, analogous to that described above for coarsening, has been the supposed impossibility of preventing spontaneous recrystallization. If the spontaneous formation of recrystallization nuclei could be prevented, even though nuclei artificially provided in various orientations were capable of growing, then the possible dependence of the rate of growth of these nuclei on their orientation could be studied.

It was recently found[11] that in high purity aluminum the spontaneous formation of recrystallization nuclei is rather rare up to deformations as high as 40 to 50 percent by rolling. Instead, essentially strain-free material is formed by migration of certain grain boundaries already present in the cold-worked material (Fig. 1). Starting with a relatively large initial grain size in order to minimize the effect of these boundary migrations, in the present experiments large grain-areas were made available for artificial nucleation. Artificial nucleation was effected by inscribing with a sharp needle one or two scratches into the surface of each large grain of the electrolytically polished portion of high purity aluminum specimens, which had been previously rolled 12 percent. After scratching, the specimens were annealed in salt bath for 1 hour at 350°C, and electrolytically etched to produce a thin oriented oxide layer, revealing variations in grain orientation.[12] Typical results are shown in Fig. 2. It is clear that in the immediate vicinity of the scratches many small grains were formed by recrystallization caused by the high local deformation. Of these small grains some served as nuclei for larger grains growing at the expense of the lightly (12 percent) deformed original grains. The relative orientation of the largest recrystallized grains with respect to the deformed matrix crystal was determined

FIG. 2. Many small recrystallized grains of various orientations at the immediate vicinity of the two scratches. Selective growth of those recrystallized grains with suitable orientation with respect to the cold-worked matrix grain. High purity aluminum initially annealed 4 hr. at 650°C, rolled 12 percent, electrolytically polished and scratched, annealed 1 hr. at 350°C. Repolished and anodically etched. The two scratches were partially removed by repolishing in order to make the numerous small grains more clearly visible. Polarized light, Biot-Klein quartz plate. Magnification 25×.

by the etch pit and optical goniometer method to an estimated accuracy of about ±1°. The results for 19 recrystallized grains found in 6 deformed matrix grains are shown in Fig. 3, with all matrix poles rotated to the same position, the "active" [111] axis being in each case in the center of the pole figure. The direction of rotation was reversed for some crystals so as to correspond in Fig. 3 to that of the other crystals. In all cases the relative orientation of the large recrystallized grains with respect to the matrix corresponded approximately to a [111] rotation of 40°. In some matrix grains two or even three [111] axes were found to be active. In Fig. 3 these are all separately rotated to the center of the pole figure.[13]

† This conclusion was reached in references 1 and 4. Kronberg and Wilson (see reference 3) attempted to explain their results in terms of oriented nucleation.

[10] C. G. Dunn, "Recrystallization textures. Symposium on cold working of metals," Trans. A.S.M., 41A (1948).

[11] P. A. Beck and P. R. Sperry, "Strain induced grain boundary migration in high purity aluminum," J. App. Phys. 21, 150 (1950).

[12] Paul Lacombe and Louis Beaujard, "Étude metallographique et cristallographique de la croissance et de la structure des pellicules d'oxydation anodique de l'aluminum," in the book published by the Comité Général D'Organisation des Industries Mecaniques, "*Etudes sur les Aspects des Pellicules D'Oxydation Anodique Formées sur L'Aluminum et ses Alliages*" (1944).

Andre Hone and E. C. Pearson, "Grain orientation in aluminum revealed by anodic film," Metal Progress 53, 363 (1948).

P. R. Sperry, "Method for studying grain boundary migration in aluminum," Trans. A.I.M.E. 188, 103 (1950).

[13] The fairly well-defined orientation relationship found here with artificial nucleation (Fig. 3) is in sharp contrast to the large scattering of orientations found by Burgers [W. G. Burgers and J. C. M. Basart, "Rekristallisation von Aluminium-Einkristallen I.," Zeits. f. Physik 51, 545 (1928)] for the recrystallized grains in slightly deformed aluminum single crystals, without artificial nucleation. In accordance with a suggestion by F. C. Frank, the University of Bristol, England, it is very likely that the scatter of orientations in Burgers' experiments was a result of lack of nuclei in orientations most favorable for growth. Such a condition could have resulted from the small extent of the deformation used. In the present experiments nuclei of a sufficient variety of orientations were produced artificially. In the recrystallization experiment with copper, described by Kronberg and Wilson (see reference 3), the lightly deformed matrix was polycrystalline, with a cube texture. The sharp recrystallization texture found by these investigators was probably a result of the availability of a large variety of nucleus orientations, due to inhomogeneous deformation at grain boundaries.

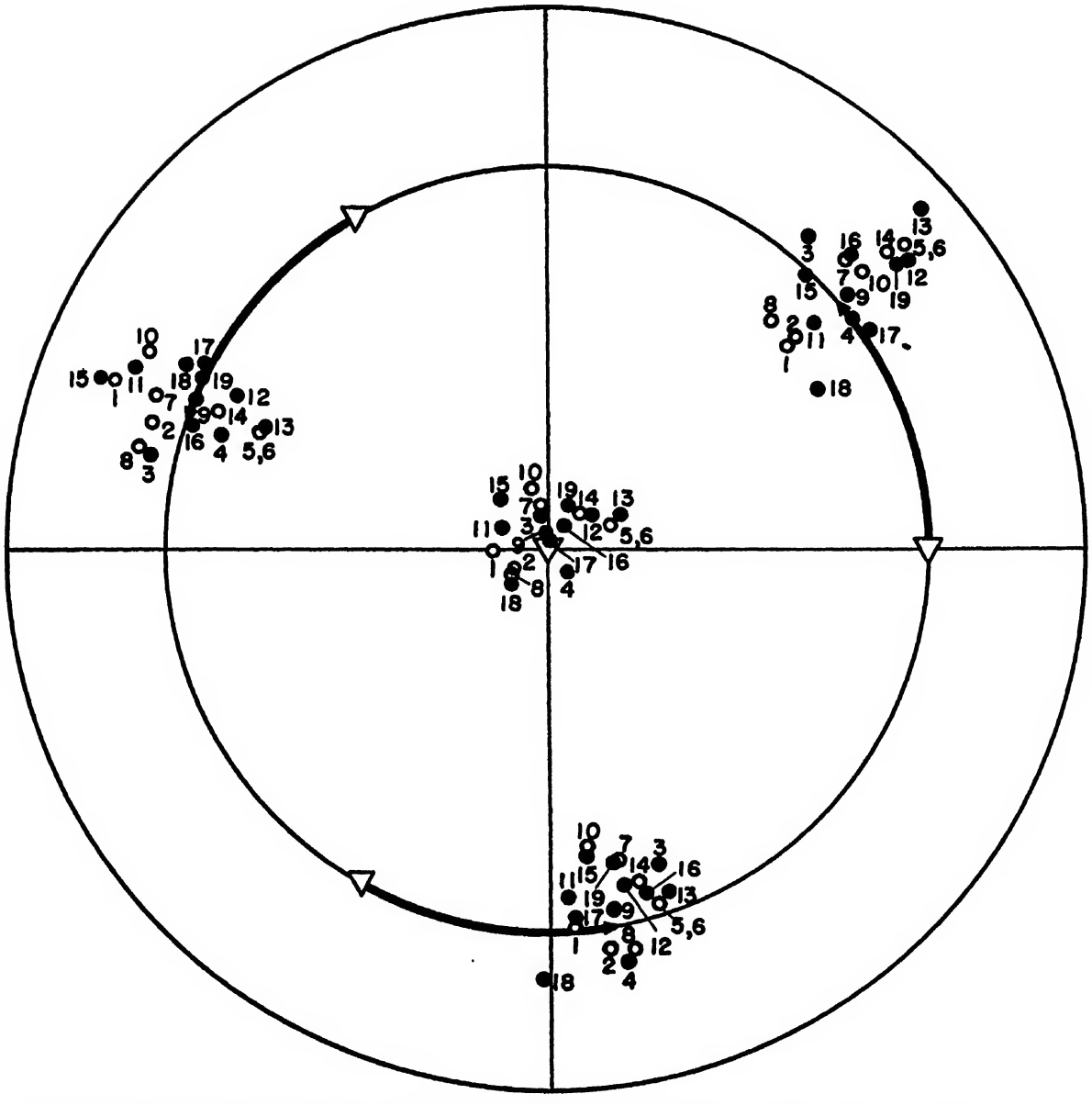


FIG. 3. (111) pole figure showing relative orientation of matrix grains △ (with the active [111] axis in the center) and of 19 largest recrystallized grains (circles). Arrow shows [111] rotation of 40°. Direction of rotation was reversed (by using mirror image of poles) for poles marked ●.

From the results shown it is evident that for recrystallization, as found previously for coarsening, the rate of growth of the artificially nucleated growing grains is highly dependent on their orientation relative to the matrix. In the recrystallization of high purity aluminum the crystal orientation which develops the highest rate of growth corresponds to a [111] rotation of about 40°–45° with respect to the matrix, as previously found for coarsening.[1] Since the same orientations were found[1] to predominate in recrystallization with spontaneous nucleation, it follows that the textures obtained in ordinary recrystallization (as well as in coarsening) can be accounted for by oriented growth, without recourse to oriented nucleation. It is interesting to note the high degree of independence of the orientation relationship between the fast-growing recrystallized grains and their matrices (Fig. 3) from variations in the orientation of the deformed matrix-grains with respect to the main directions of the rolling operation (Fig. 4). Another interesting observation may be made by inspection of Fig. 2. In certain directions the new grains grow much faster than in others; the rate of growth of the recrystallized grains is quite anisotropic. It appears that (111) faces tend to develop, suggesting that the rate of migration of such faces is lower than that of others.

Considering the controlling importance of the orien-

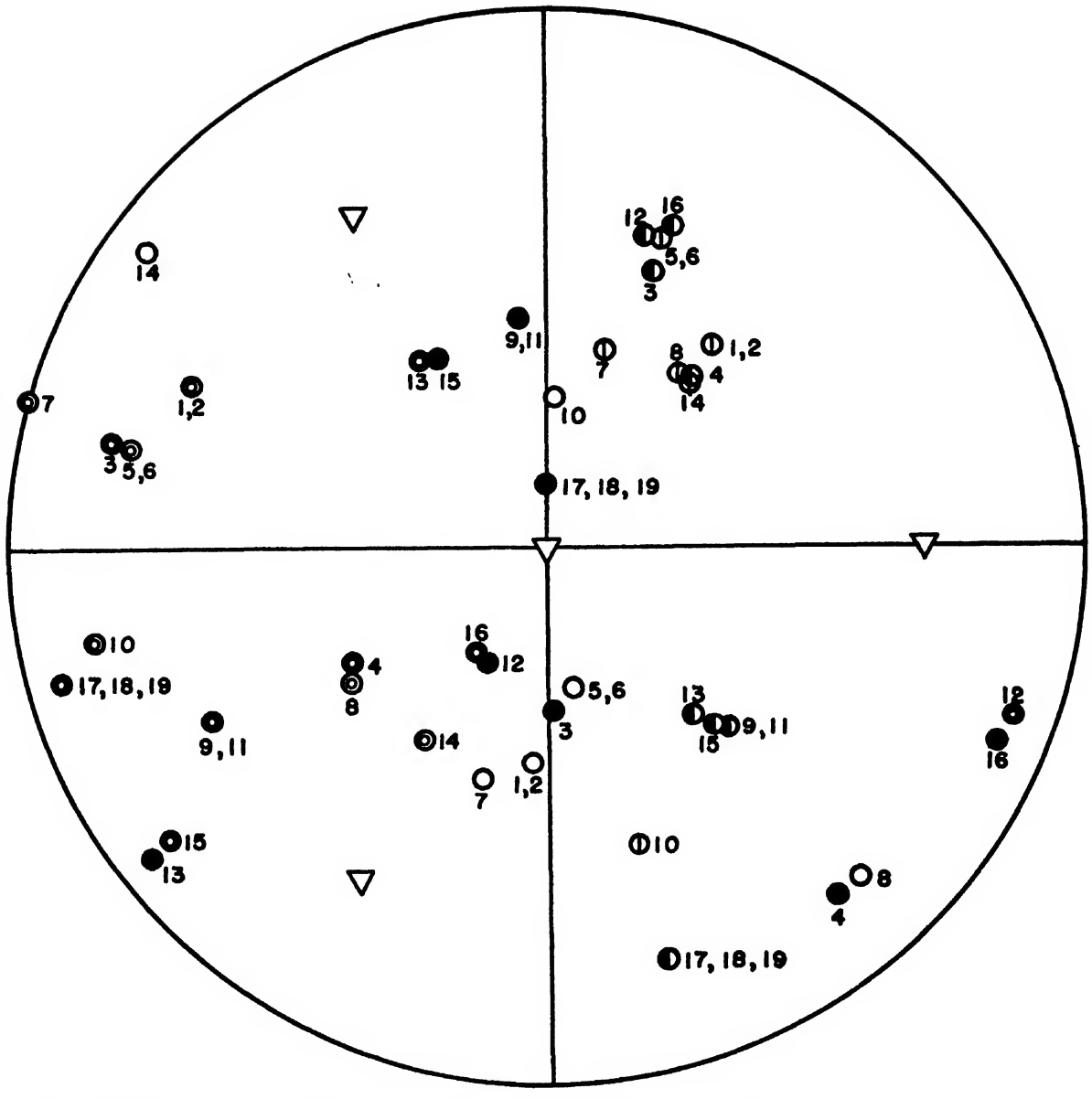


FIG. 4. Pole figure showing orientation of matrix grains △, with respect to the direction of rolling ○, the normal direction ● and the cross direction ⊕ (active [111] axis in the center). The direction of rotation was reversed where circles are partially or completely filled.

tation-dependence of the rate of grain boundary migration, one would like to form a picture of this effect in terms of fundamental physical processes. In this connection, it is worth noting that the strain-free grains formed in cold-worked and annealed high purity aluminum in the manner shown in Fig. 1, are always growing at the expense of a strained grain of different orientation.[11] In observing a large number of such grains, they were *never found to grow at the expense of the strained grains of nearly the same orientation* with which they are in contact.[14] That the absence of migration of the boundary between a strained and a strain-free grain of nearly the same orientation is not due to

[14] It is interesting that, on the other hand, the sub-boundaries observed by P. Lacombe and L. Beaujard ["Les imperfections de structure des cristaux uniques d'aluminium pur," Revue de Metallurgie 45, 317 (1948)] in polygonized aluminum are apparently highly mobile [A. Guinier and P. Lacombe, "L'état polygonisé, du cristal metallique," Métaux et Corrosion (1948), No. 212], even though they are located between lattice regions of nearly the same orientation. Possibly, this high mobility is connected with the fact that such a sub-boundary may be considered as a set of parallel line dislocations; the mechanism of boundary migration here is probably the one proposed by W. G. Burgers ["Recovery and recrystallization viewed as processes of dissolution and movement of dislocations," Proc. Koninklÿke Nederlandsche Akademie van Wetenschappen, 50, 452 (1947)]. On the other hand, in the more general case, where two sets of intersecting edge dislocations are necessary to describe the boundary, it is more

FIG. 5. Strain induced boundary migration between grains *C* and *B* and between grains *A* and *B*. Boundary positions before and after the last anneal are marked 1 and 2. Strain-free area swept by moving boundary between *C* and *B* has orientation of *B*; it is growing into *C*. Strain-free area swept by migrating boundary between *A* and *B* has same orientation as *A*, and it is growing into *B*. High purity aluminum annealed at 650°C for 4 hr., rolled 12 percent, electrolytically polished, annealed 4 hr. at 350°C and 1 min. at 400°C, repolished mechanically and electrolytically and etched anodically. Oblique illumination. Magnification 75×.

a lack of driving energy[15] is obvious from the fact that the same strained grain is often seen being absorbed by a strain-free grain of a different orientation (Fig. 5). The surface energy connected with the migrating boundary (between two grains of different orientation) is undoubtedly larger than that connected with the stationary boundary (between two grains of practically the same orientation). Hence the energy required to increase the surface area of the boundary actually migrating is larger than the energy that would be required to expand the boundary which is in fact stationary; it so happens that strain induced boundary migration occurs just where the surface energy conditions are unfavorable for it. Clearly, then, the deciding factor must be something other than surface energy. It has been pointed out by Kronberg and Wilson[3] that the experimentally found relative orientations are near ones at which a certain percentage (such as 7 percent) of the atomic positions in the two lattices (that of the growing crystal and of the matrix) coincide, whereas generally in other orientations such coincidences are very rare. That an orientation relationship involving a large number of coincidences is not necessarily favorable for fast boundary movement is clear from the fact described above, that with nearly 100 percent coincidence the boundary is practically immobile. The fact that twin boundaries are particularly stable also points in the same direction. An exceptionally convincing demonstration of the lack of mobility of twin boundaries in the face of considerable driving forces was recently given by C. S. Smith.[16]

The experimental facts may be summarized by stating that with relative lattice orientations corresponding to a [111] rotation of 0° or 60° (twin orientation) the mobility of the grain boundaries is very low, and that with relative lattice orientations between these two, corresponding to [111] rotations of about 30°–45°, the mobility is the highest. Let us assume that grain boundary migration is a result of atoms leaving unfavorable positions in one crystal lattice, diffusing along the grain boundary and finally attaching themselves at some particularly favorable location to the crystal on the other side of the boundary. The necessity of *self-diffusion along the boundary* is an essential feature of the assumed mechanism of grain boundary migration. The effect of the relative orientation of the two crystals on the "mobility" of the boundary between them must be a result of its influence on the rate of transfer of atoms across the boundary. In view of the above mechanism of boundary migration, the most probable assumption as to the manner in which such an influence may be exercised is that the relative lattice orientation has a strong effect on the rate of self-diffusion along the grain boundary. It can be readily visualized for instance that at the boundary between two crystals of nearly the same orientation, or at a twin boundary, there is relatively little disturbance; the atoms are more tightly held in their positions, hence the rate of their diffusion is lower than along a boundary with more disorientation. Such an effect has apparently not as yet been found in diffusion studies, but its existence could be relatively easily ascertained by comparing the rate of self-diffusion along grain boundaries in a specimen of a single-orientation texture with that in a specimen of random orientation, other factors, such as grain size, being held constant.

CONCLUSION

(1) Evidence is presented to show that in slightly deformed high purity aluminum crystals the rate of growth of recrystallized grains of various orientations varies over wide limits.

(2) Those recrystallized grains which grow at the

likely that grain boundary migration takes place by individual atomic migration, as described further below.

[15] The fact that in fine grained high purity aluminum gradual grain growth is impeded in the presence of a strong single orientation texture [P. A. Beck and P. R. Sperry, "Effect of recrystallization texture on grain growth," Trans. A.I.M.E. 185, 240 (1949)] is probably closely related to the described effect. However, in this case the situation is not nearly as clear-cut as in the phenomenon just discussed, since the driving energy (in this case grain boundary surface energy) as well as the mobility becomes very small. The same applies to the remarkable stability of a sharply defined "cube texture" in copper (see reference 3), and probably also to the small "insular" grains left unabsorbed by a surrounding recrystallized grain of nearly the same orientation [P. Lacombe and Aurel Berghazen, "Relations d'orientation entre monocristaux métallique de recristallization et petits cristaux inclus," Métaux et Corrosion (1949), No. 281].

[16] Discussion by C. S. Smith of the paper by Hibbard, Liu, and Reiter, "Annealing twins in copper and 70–30 alpha-brass," Trans. A.I.M.E. 185, 635 (1949).

highest rate have relative orientations with respect to the matrix corresponding approximately to a 40° rotation around a common [111] axis.

(3) Since the orientation with the highest rate of growth closely approximates that found previously for the recrystallization texture of 70 to 80 percent rolled or compressed aluminum single crystals, it is concluded that the recrystallization texture can be accounted for by the orientation dependence of the rate of grain boundary migration. (Similar conclusion was reached earlier (see reference 1) regarding the coarsening texture.)

(4) The hypothesis is advanced that the orientation dependence of the rate of grain boundary migration is brought about by the dependence of the rate of self-diffusion along grain boundaries upon the relative orientation of the two grains meeting at the boundary. The assumed diffusion effect could be tested independently from its connection with grain boundary migration.

Measurement of the Thermal Diffusivity of a Soil by the Use of a Heat Pump

E. B. PENROD
Department of Mechanical Engineering, University of Kentucky, Lexington, Kentucky
(Received October 13, 1949)

An earth heat pump was set in operation to abstract heat from soil assumed to be in thermal equilibrium. Temperature-time graphs of the antifreeze and of the soil at distances of six, twelve, and eighteen inches below the center of the antifreeze line were plotted from experimental data. From these graphs data are obtained from which the thermal diffusivity of the soil is calculated. The soil was analyzed and found to fall in the general Casagrande classification of lean clay (*CL*) with low plasticity.

INTRODUCTION

IN designing the ground coil or earth heat exchanger for a heat pump system that uses the earth as its source of heat, the thermal properties of the soil and its general classification should be known.* When the earth heat pump[1] is operating on the heating cycle, for example, heat is absorbed from the soil by the cold fluid circulating through the earth heat exchanger. The time rate of heat flow through the earth depends partly on its moisture content which in turn depends on the grain size of the soil.

An original method is used to determine the thermal diffusivity of soils which may be applicable to geological and geophysical studies. The soil was analyzed and classified in addition to measuring its thermal diffusivity.

THE METHOD

After the earth heat exchanger had been buried in the soil a transient temperature distribution was created in the ground. This was accomplished by circulating a cold antifreeze solution continuously through the earth heat exchanger (Fig. 1) and the evaporator coils of the heat pump system. A section of the earth heat exchanger at station 3 (Fig. 2) was considered as an infinite line heat source when the heat pump was operating on the heating cycle.

It was assumed that the initial temperature of the soil is a function $f(r)$ of the distance from the antifreeze line and that the temperature at any time τ is a function $t(r, \tau)$. The differential equation for transient heat flow[2] in cylindrical coordinates is

$$\frac{\partial t}{\partial \tau}=\alpha\left(\frac{\partial^2 t}{\partial r^2}+\frac{1}{r}\frac{\partial t}{\partial r}\right). \tag{1}$$

Using finite differences[3] instead of differentials, Eq. (1) transforms to

$$\alpha=\frac{\Delta t/\Delta \tau}{(\Delta^2 t/\Delta r^2)+(1/r)(\Delta t/\Delta r)}, \tag{2}$$

where α is the thermal diffusivity of the soil.

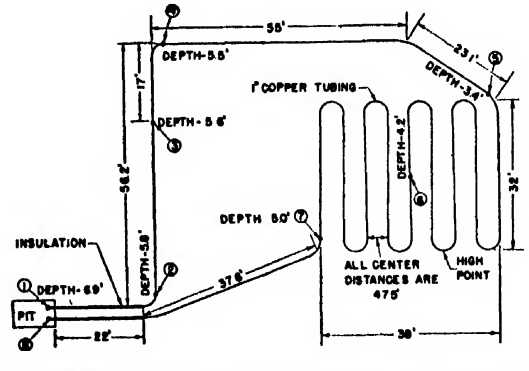


FIG. 1. Schematic diagram of the earth heat exchanger. The portion at station 3 was considered as an infinite line heat source.

* The Kentucky Utilities made the research possible by giving a Marvair heat pump and financial grant-in-aid to the University of Kentucky which supplemented the funds allocated by the Engineering Experiment Station of the University.

[1] E. B. Penrod, Am. Scientist 35, 506 (1947).

[2] R. V. Churchill, *Fourier Series and Boundary Value Problems* (McGraw-Hill Book Company, Inc., New York, 1941), p. 165.

[3] Max Jakob, *Heat Transfer* (John Wiley and Sons, Inc., New York, 1949), Vol. I, p. 392.

TABLE I. Data for calculating $\Delta t/\Delta\tau$ and $\Delta t/\Delta r$.

Curve	Distance below center of antifreeze line, r ft.	Thermo-couple No.	Temperatures at noon, March 19, 1949, °F	Change in temperature from March 16 to 22, 1949, Δt°F	Time interval, $\Delta\tau$ hours
A	0.0	3	30.6	− 9.7	144
B	0.5	12	44.5	−10.9	144
C	1.0	13	48.6	− 9.0	144
D	1.5	14	50.1	− 5.0	144

The experiment was started on March 16, 1949 when temperatures of the antifreeze and soil were recorded at regular intervals for two days before the heat pump was put in operation. At 9:00 A.M., March 18, the plant was set in operation so that heat was absorbed from the soil as the antifreeze was circulated through the earth heat exchanger. While the heat pump was in operation temperatures were taken at regular intervals in the center of the antifreeze line (Fig. 2) at station 3 and in the soil below it at distances of six, twelve, and eighteen inches respectively. These data are shown graphically in Fig. 3.

Data for calculating the slope, $\Delta t/\Delta\tau$, of curves *A*, *B*, *C*, *D* at distances $r=0.0$, 0.5, 1.0, and 1.5 feet below the center of the antifreeze line are given in the last two columns in Table I. Curve *I* (Fig. 4) shows the variation of the slope with respect to distance below the center of the antifreeze line.

The temperature gradients, or first differences, were calculated from the data given in the fourth column of

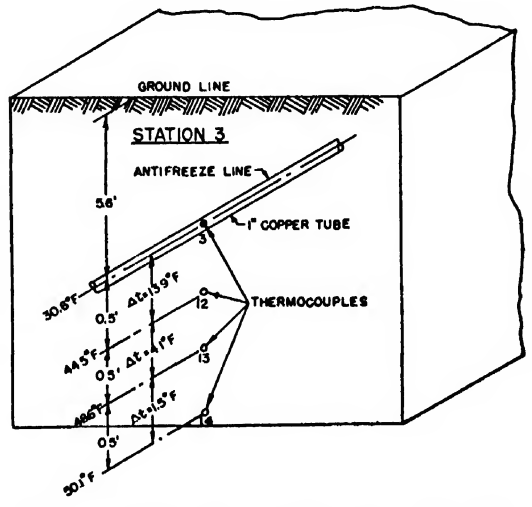


FIG. 2. Schematic diagram showing the location of the thermocouple beads at station 3.

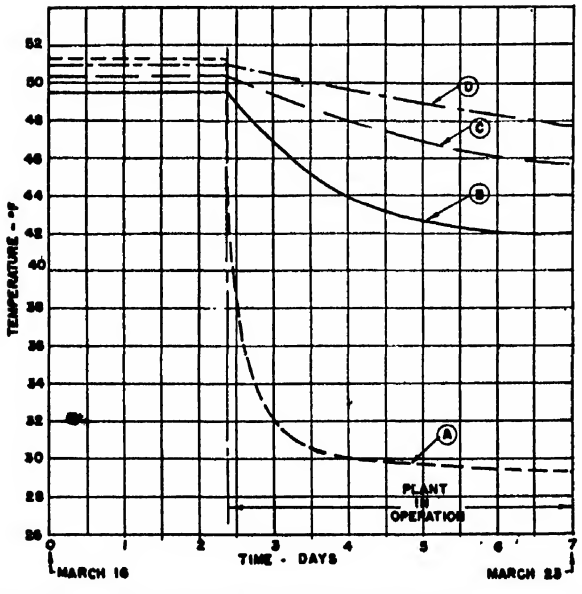


FIG. 3. Graphs showing the variation of temperature with respect to time at station 3.

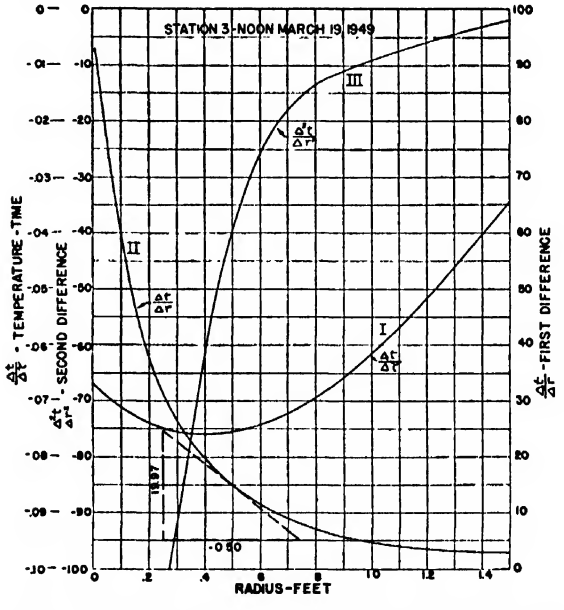


FIG. 4. Graphs used to determine the thermal diffusivity of the soil at station 3. Curve *I* was obtained from curves *A*, *B*, *C*, and *D* of Fig. 3 for noon, March 19. Curves *II* and *III* show the variation of the first and second differences with respect to the distance below the center of the antifreeze line.

Table I, and are listed in the third column of Table II. Curve *II* (Fig. 4) was obtained by coordinating values of the temperature gradient and distance from the center of the antifreeze line. The second differences, $(\Delta^2 t/\Delta r^2)$, were obtained by finding the slopes of curve *II* (Fig. 4) at various distances from the center of the antifreeze line. These results are given in the fourth column of Table II.

By substituting data given in Table II, in Eq. (2) values of the thermal diffiusivity of the soil were obtained.

SOIL CLASSIFICATION

It is necessary to know the thermal properties of the soil as well as its general classification before an efficient

TABLE II. Calculated data and results.

Distance below center of antifreeze line, r ft.	Noon, March 19, 1949 $\Delta t/\Delta r$, °F/hour	$\Delta t/\Delta r$, °F/ft.	$\Delta^2 t/\Delta r^2$, °F/ft.²	Thermal diffusivity α, ft.²/hour
0.00	−0.0673			
0.25	−0.0752*	27.8	−115.0	0.0198
0.50	−0.0757	18.0	− 39.94	0.0192
0.75	−0.0715*	8.2	− 15.0	0.0176
1.00	−0.0625	5.6	− 9.0	0.0184
1.25	−0.0490*	3.0	− 5.0	0.0189
1.50	−0.0347			
			Average	0.019

* Values taken from curve *I*, Fig. 4.

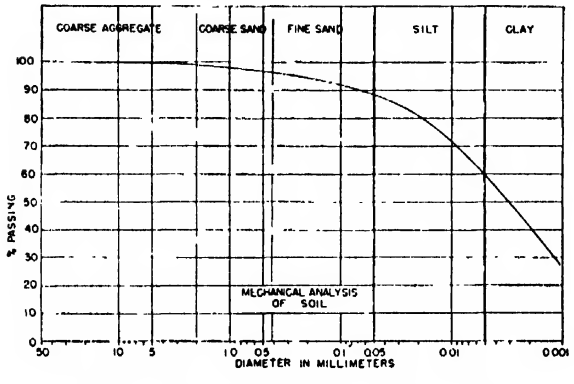


FIG. 5. Graph showing grain size of the soil in the vicinity of station 3.

TABLE III. Physical test data of the sample of soil analyzed.

Liquid Limit

Dish No.	No. of blows	Wt. of dish and wet soil	Wt. of dish and dry soil	Wt. of dish	Wt. of water	Wt. of dry soil	% of water	Liquid limit
66	29	9.01	8.00	5.56	1.01	2.44	41.4	43.4
29	24	9.98	8.63	5.56	1.35	3.07	44.0	

Plastic Limit and Plasticity Index

Dish No.	Wt. of dish and wet soil	Wt. of dish and dry soil	Wt. of dish	Wt. of water	Wt. of dry soil	% of water	Plastic limit	Plasticity index
72	6.26	6.10	5.55	0.16	0.55	29.1	27.3	16.1
142	5.94	5.84	5.45	0.10	0.39	25.6		

Casagrande classification—*CL*.

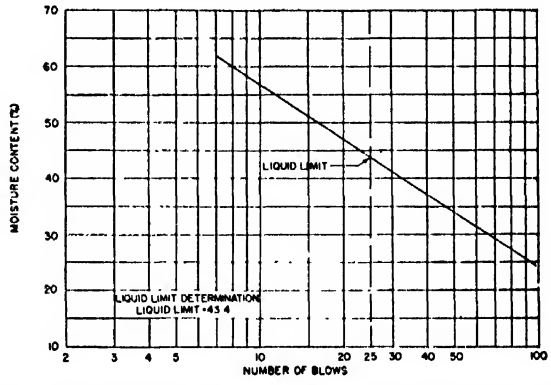


FIG. 6. Graph showing the moisture content of the soil in the vicinity of station 3.

earth heat exchanger can be designed. In order to determine the characteristics of the soil in which the earth heat exchanger was buried, a sample of the soil was taken at a depth of about 5 feet and 6 inches in the vicinity of station 3 at the time of the trench excavation.† Table III contains physical test data on the sample of soil, while results of the test for grain size distribution and plasticity characteristics are shown in Figs. 5 and 6. About 60 percent of its grains are in the size generally termed clay, and less than 30 percent of the sample is silt. The liquid limit of the sample is 43.4 percent and its corresponding plasticity index is 16.1 percent. The sample of soil falls in the general Casagrande classification of lean clay (*CL*) with low plasticity characteristics.[4]

By core sampling, the density of the soil in the vicinity of station 3 was found to be 120 pounds per cubic foot as received and 103.4 pounds per cubic foot after being dried to constant weight.

CONCLUSIONS

The average value of the thermal diffusivity of the soil was found to be 0.019 ft.² hr.$^{-1}$ (0.0049 cm² sec.$^{-1}$) for the soil at a density of 120 pounds per cubic foot. The soil falls in the general Casagrande classification of lean clay (*CL*) with a low plasticity.

† The soil was analyzed by Messrs. L. E. Gregg and N. B. Drake of the Highway Materials Research Laboratory, Department of Highways of the Commonwealth of Kentucky.

[4] "Classification and Identification of Soils," A.S.C.E. Trans. 113, 901 (1948).

I. Indoor Measurement of Microwave Antenna Radiation Patterns by Means of a Metal Lens

G. A. Woonton, R. B. Borts, and J. A. Carruthers
Department of Physics, McGill University, Montreal, P.Q.
(Received October 19, 1949)

The *E*-plane radiation patterns of two horn radiators have been measured in the laboratory with a metal lens by the same method that is used in optical diffraction pattern measurements. Comparison of patterns measured outdoors, by standard methods, with the lens measurements shows little deterioration in precision over the angular interval between ±15 degrees, for horns of aperture width 32 cm, measured by a lens of aperture 110 cm, at a wave-length of 3.2 cm. It was found that a stepped lens was unsatisfactory for this purpose, possibly because of diffraction caused by the steps.

INTRODUCTION

THE radiation pattern of an antenna is a statement of the power that the antenna abstracts from a plane electromagnetic field, expressed as a function of its angular position in the field. The pattern usually takes the form of a graph between power, relative to maximum received power, and angle of rotation measured from the point of maximum reception. Microwave antenna patterns are measured by rotating the antenna on a turntable, in a field set up by a transmitter some tens, hundreds or thousands of feet away. It has been shown by Cutler, King, and Kock[1] that the pattern produced in this way, approximates the true pattern to the extent that the spherical wave front from the transmitter approximates a plane wave in both phase and amplitude. It follows from Eq. (11) of their paper that, as the wave-length is decreased, it becomes increasingly more difficult to find a suitable, clear site for antenna pattern measurements, or conversely, that for a given site there is a wave-length limit below which the precision of measurement deteriorates.

Experiments have shown that, under favorable conditions, antenna patterns can be measured in the laboratory, with the help of a lens, by the same method as that used in optical diffraction measurements. Since both the precision and the convenience of making measurements of this type may be expected to increase with decreasing wave-length, it appears, that in these optical-type measurements there may be found a technique which will supplement standard radio methods in the wave-length range where a small outdoor site tends to become useless. The method is as follows: The spherical wave front from a point source is made plane for pattern measurement, by a metal lens, to an approximation that is determined by incidental diffraction from the aperture of the lens. Subject to restrictions on wave-length, on the sizes of the antenna and the lens, and on the angle of rotation, errors in the laboratory measurement due to incidental diffraction can be calculated and the calculations can be used to correct the measured pattern to the true plane wave pattern. The theory of these calculations, for linear radiators, has been discussed by G. A. Woonton.[2]

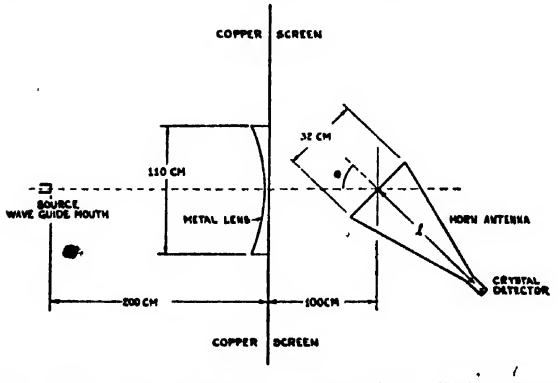


FIG. 1. Plan view of laboratory set-up for radiation pattern measurements at a wave-length of 3.2 cm.

EXPERIMENTAL METHODS AND RESULTS

Figure 1 is a schematic plan view of the measuring apparatus. Figure 2 is a photograph giving some of the details of the equipment. All measurements were made at a wave-length of 3.2 cm.

The source of radiation was the open end of a piece of 1 in.×½ in. wave guide which had been terminated with a plane flange. The open end was fed with 3.2 cm energy, through a wave meter and a calibrated antenuator section, from a Type 723A/B Klystron. The open end of the guide was placed about five feet above the laboratory floor, in line with the center point of the metal lens and at the focal point which was 200 cm behind the back surface of the lens.

Two metal lenses were used in these experiments, both of which were designed, from formulas given by Kock,[3] to have a focal length of 200 cm. The plate spacing for both lenses was chosen to be 2 cm which resulted in an effective refractive index of 0.6, at a wave-length of 3.2 cm. The first of these was a stepped lens of square aperture, 114 cm to the side, and was designed to focus in all planes. The plate spacing was maintained by five columns of metal spacers. It was found that antenna patterns, measured with this lens, exhibited unpredictable divergences from the true

[1] Cutler, King, and Kock, Proc. I.R.E. 35, 1462 (1947).

[2] G. A. Woonton (to be published).

[3] W. E. Kock, Proc. I.R.E. 34, 828 (1946).

radiation pattern and, after a series of modifications, the lens was discarded. The evidence is not conclusive, but it is believed that the final failure of this lens is associated with diffraction from its steps. During the tests with this lens it was found that the metal plate-spacers, which were $\frac{3}{16}$-in. metal cylinders, contributed unwanted diffraction, as did substitute Bakelite spacers. This trouble was eliminated by spacing the plates with acetate foam blocks. The performance of the lens was critically sensitive to the transverse position of the source. The condition was remedied by shielding most of the lens aperture so that the radiation passed through a rectangular slit 110 cm long by 22 cm high, with its long dimension in the plane of measurement.

The second lens was designed for the measurement of linear radiators only, and has been found satisfactory; the measurements which appear in Figs. 3, 4, and 5 were made with it. The lens contour is an elliptic cylinder which focuses in the *E*-plane. The aperture is a rectangle 114 cm in the *E*-plane by 20.6 cm in the *H*-plane. The plates are spaced by slabs of Styrofoam which completely fill the inter-plate space. The lens was mounted in a wooden frame with its center plate parallel to and five feet above the floor. Energy from the source was prevented from reaching the antenna, by paths other than through the lens, by means of a copper screen approximately nine feet square around the lens. The front, uncurved surface of the lens fitted snugly against the edge of an aperture (110 cm by 20.6 cm) which was cut in this screen; the edges of the aperture were given sharp definition by soldering the copper screen around a frame made of formed brass rod.

The correct position of the lens and source and the general adequacy of the set-up were determined by comparing the measured field in front of the aperture with the predicted diffraction field; simple optical theory, applied to the case of a plane wave, of length 3.2 cm incident on an aperture 110 cm by 20.6 cm was used in the calculation. Figure 3 shows that, by careful adjustment, the measured field can be made to approximate the predicted field. In the course of these measurements, it was learned that strong reflections from the windows and the walls could be minimized by setting the lens at an angle to the reflecting surface. The field measurements were made with a dipole and crystal detector mounted on an optical bench. The auxilliary apparatus was the same as that used for antenna measurements.

The *E*-plane patterns of the two horn radiators have been measured by this lens method; Figs. 4 and 5 give the results of these measurements. The field in which the horns were measured was square-wave modulated at six hundred cycles per second by modulating the reflector voltage of the klystron. A crystal detector attached to the apex of the horn converted the modulated received field into an audio voltage, which was amplified by a tuned audio amplifier and measured by a diode and direct current microammeter. The system was calibrated, before the measurements, by plotting the direct current output from the diode against the attenuation setting of a precision, calibrated, vane-type attenuator, which had been interposed between the klystron and the wave guide mouth. The antenna pattern was constructed by recording relative power in decibels (as determined by the microammeter and this graph) against the angular position of the horn in the field.

The copper horns on which experimental measurements were made were flared linearly in both the *E*- and *H*-planes from the end of a piece of standard, laboratory, *X*-band wave guide to a square aperture 32 cm to the side. Figure 4 presents pattern data on one of these horns, which was approximately 50 cm long; Fig. 5 presents data on the other which was about 100 cm long. Results from three different experiments are compared on each of these graphs: the solid line is the horn pattern, measured at an outdoor site, over a measuring path of about 225 feet, following standard procedure; it is considered to be a close approximation to the true antenna pattern. The dashed line is the mutilated distant pattern; a rectangular aperture, of the same dimensions and orientation as that used in the laboratory with the lens, was cut in the center of a large copper screen. The screen and its aperture were erected, at the outdoor site, one meter in front of the horn, between the horn and the distant measuring point 225 feet away. The horn pattern was then measured, following standard procedure. The pattern of the horn, as measured in the lens field in the laboratory, is indicated by crosses.

CONCLUSION

The radiation pattern of a linear antenna can be measured in the laboratory by optical methods, to a precision that is set only by incidental diffraction from the aperture of the lens. If the lens aperture is several times larger than the antenna aperture, the contribution

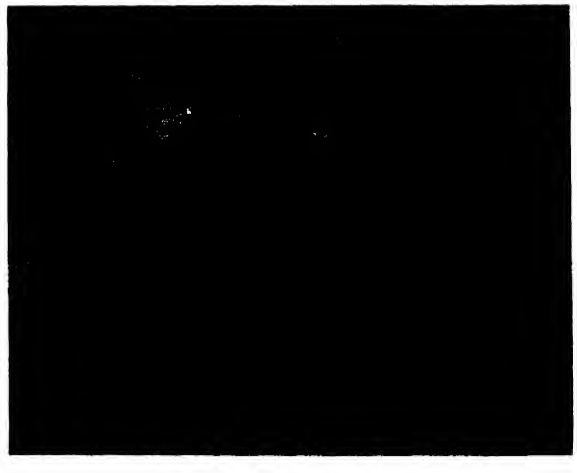


FIG. 2. Equipment used in laboratory measurement of antenna patterns.

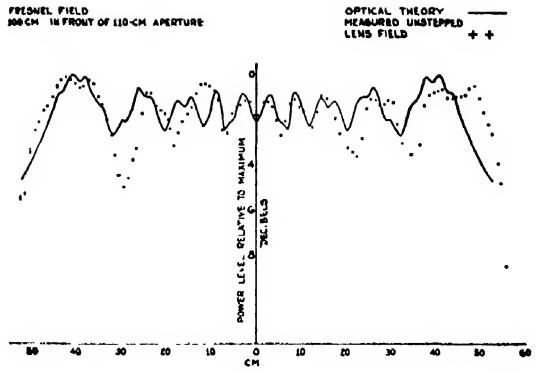


FIG. 3. Comparison of measured and predicted Fresnel field at a distance of 100 cm in front of the lens.

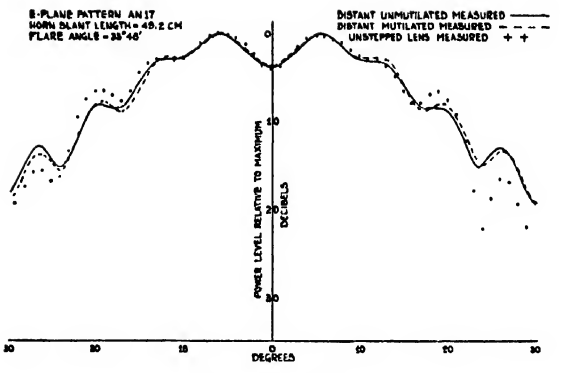


FIG. 4. Comparison of the true radiation pattern, the mutilated distant radiation pattern, and the radiation pattern measured through the lens for the *E*-plane of the 50-cm horn.

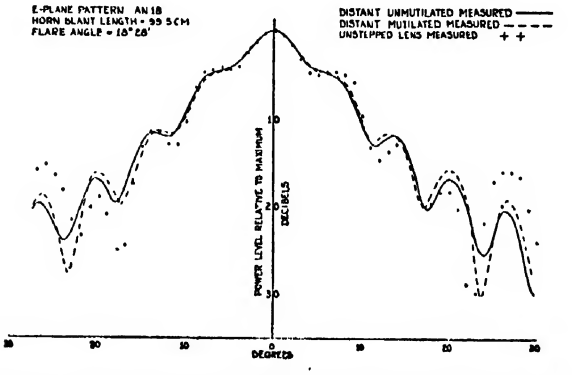


FIG. 5. Comparison of the true radiation pattern, the mutilated distant radiation pattern and the radiation pattern measured through the lens for the *E*-plane of the 100-cm horn.

of incidental diffraction to the center region of the pattern is negligible. Good agreement between patterns measured by the lens method and by the standard outdoor technique was obtained over the angular interval $-15° \leq \alpha \leq 15°$ for a lens aperture of 110 cm, an antenna aperture of 32 cm, and at a wave-length of 3.2 cm; beyond $\pm 15°$ the measured positions of the side lobes are approximately correct, but the relative amplitudes depart from those found by outdoor measurements. The effect of room reflections on these measurements can be minimized by proper location of the apparatus and when so minimized, reflections are not a serious source of error. A stepped lens has been found to be unsatisfactory in this application, probably because of diffraction caused by the steps and in general it has been found necessary to take more care in the design and construction of a lens for measuring purposes than has seemed essential with those to be used as antennas. Although it was found necessary to limit these measurements to the *E*-plane of the horn radiators, it is believed that there is no reason, other than physical size, why the method cannot be applied to all planes of an antenna of any kind.

ACKNOWLEDGMENT

This research was made possible by a grant from the Defence Research Board of the Department of National Defence, Ottawa, Canada. The stepped lens was built and some of the first experiments on it were performed by Messrs. J. H. Blackwell and H. M. Dutton with the senior writer, at the University of Western Ontario. Funds for the construction of the stepped lens and other apparatus used at that time were supplied from a grant made by the National Research Council, Ottawa, Canada. One of us (R.B.B.) is indebted to the National Research Council for a Bursary during the winter term of 1948–49.

Load Induced X-Ray Line Broadening in Nylon Filaments

JOHN KAUFFMAN* AND WALLER GEORGE*
Naval Research Laboratory, Washington, D. C.
(Received October 21, 1949)

The load induced microscopic structural changes in highly oriented Nylon filaments were followed by changes in the x-ray fiber diffraction pattern as a function of the resulting strain. Observations made included (1) changes in the dimensions of the fiber spots, and (2) a decrease in ratio of the intensity of amorphous to crystalline x-ray scattering. These new results for Type 200 du Pont Nylon indicate a change in the micro deformation mechanism at a strain corresponding to the inflection of the load-strain diagram (14 percent elongation). The mechanism for smaller plastic strains appears to involve a rotation of the crystallized domains into alignment with the fiber axis, accompanied by a small over-all additional crystallization. For strains larger than the inflectional strain, the mechanism seems to involve a progressive reduction in the domain size, and a reduction in alignment with the fiber axis. Both effects appear to saturate in the vicinity of the rupture strain. These results are discussed in terms of load induced effects found in metal single and polycrystals. It is suggested that the qualitative features observed may be representative of a wide class of load or strain induced micro deformation reactions in the solid state.

INTRODUCTION

THIS paper** describes a portion of a general research directed toward an understanding of the micro mechanisms of plastic flowing in crystalline high polymers. X-ray diffraction was the tool used for the study, and the material chosen was a strong, highly oriented, crystalline Nylon fiber obtained from E. I. du Pont de Nemours, Inc., denoted as Type 200. The fibers were received as 100 denier monofilaments.

The fine structure of oriented Nylon approaches the state found in polycrystalline metals, but in the case of crystallized polyamides, the crystallites are composed of long, complex molecules instead of atoms. These linear molecules exhibit a high degree of geometric organization within the region of the crystallite. The boundaries between crystallites, while imperfect and ill-defined, consist of what is essentially an "amorphous" material composed of molecules in a disordered state, which may be pictured as coiled and interlaced without geometric simplicity.

The microstructural state of 100 denier filaments of Type 200 is indicated by the x-ray fiber patterns as shown in Fig. 1. They are qualitatively similar to those obtained from polyamide fibers by others.[1] These were obtained using Ni filtered Cu$K\alpha$-radiation at room temperature with film specimen distance of 6 cm. The spots and arcs in the pattern result from periodicities in the crystalline material which scatter the radiation coherently. The disordered intercrystalline material produces the poorly defined amorphous halo at about the reflecting angle of the two very prominent equatorial spots. It is possible to obtain an estimate of the ratio of amorphous to crystalline material present in the specimen by comparing the intensity of the crystalline scattering to the amorphous scattering.[2]

EXPERIMENTAL PROCEDURE

The structure of the material as it is indicated by diffraction patterns was followed as a function of the amount of plastic strain from the initial condition of the specimen through tensile strains up to the rupture point, for strains resulting from a constant time rate of load application of 2.6 grams/denier/minute.

The specimens were in the form of a bundle of 25 monofilaments, and were deformed in tension in a commercial Scott IP-4 inclined plane testing machine.

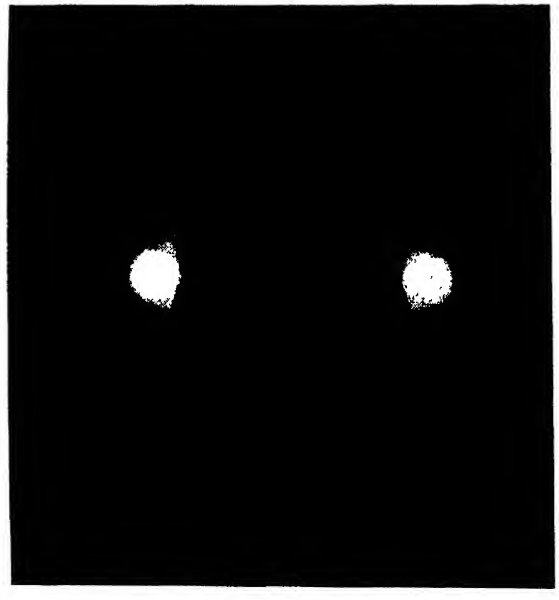


FIG. 1. Typical fiber pattern from Nylon bundle as received. Exposure *ca.* 1 hour. Film-specimen distance 4 cm Cu$K\alpha$-radiation.

* Members of the Staff, Mechanics Division.

** The new results reported in this paper comprise a portion of a thesis submitted by John W. Kauffman to the graduate faculty of the University of Maryland in partial fulfillment of the requirements for the M.S. degree (June, 1949). In abbreviated form, the results were discussed at the Cambridge meeting of the American Physical Society [Phys. Rev. 76, 467 (1949)].

[1] I. Fankuchen and H. Mark, J. App. Phys. 15, 364 (1944).

[2] W. Weidinger and P. H. Hermans, J. App. Phys. 19, 491 (1948).

The machine used was equipped with "drum"-type specimen clamps which had been highly polished to reduce the friction at the fiber-drum interface. As the specimens reached various strains, the loading was abruptly terminated, the specimen clamped at a constant strain, removed from the testing machine, and the diffraction pattern immediately taken. In order to detect effects which might have an appreciable time dependence, additional patterns were taken after the specimen had been under a constant strain for several days; however, no such effects were observed for these times. Hence, one may consider long-time relaxation effects to be negligible for these experiments and therefore it is possible to relate the structural changes shown by the diffraction pattern to the points on the load-elongation curve at which the loading of the specimen was terminated. The time elapsed from the termination of the loading cycle until the beginning of radiation was approximately thirty minutes and held essentially constant for all experiments. This time interval apparently is sufficiently long to allow completion of short-time relaxation processes in all cases. This latter point is demonstrated in Figs. 2a and 2b, which show patterns obtained within one hour after the termination of the loading epoch and forty-eight hours, respectively. The strain used in this case was almost the strain at rupture, 26 percent.

Photometer traces were taken of the spots, arcs and halo of the diffraction patterns. A full interpretation of these patterns involves tedious and laborious procedures. The present paper is concerned with their more direct and obvious features which are pertinent to a plastic deformation study. In all cases attention was confined to relative changes in the diffraction patterns which could be considered to result from plastic flowing.

RESULTS

With deformation it was observed that the diffraction pattern showed (1) changes in the dimensions of the fiber spots and (2) a decrease in the intensity of the amorphous relative to crystalline scattering.

Dimensional changes were observed in the two most prominent equatorial spots and these were of two types, which are designated as angular and radial broadening. The annular width at half-maximum intensity in a direction radially from the center of the pattern was taken as the radial broadness. It has been used by many investigators[3] for determining particle size or for detecting inhomogeneous lattice strains. One can show that the radial broadness of a diffraction spot may result from one or both of two sources: (1) Diffraction from small crystallographically ordered domains which is a function of the domain size, and (2) Diffraction from lattice planes with irregularly varying parameter possibly associated with the existence of internal strains which vary from point to point in the lattice. Simple theories of both types of radial broadness effects are essentially the same, and lead to relations for the dependence of the broadness on the Bragg angle and wave-length of the incident radiation which are identical to a multiplicative constant.[4] However, it will be seen that the data reported below lend themselves more readily to the particle size interpretation.

The angular broadness is defined as the width of the diffraction spot or line at half-maximum intensity taken in a direction at right angles to the radial broadness measurement; that is, the photometer trace is made around the pattern at a constant radius from the center. The angular broadness is taken as a measure of the degree of preferred orientation of the crystallites

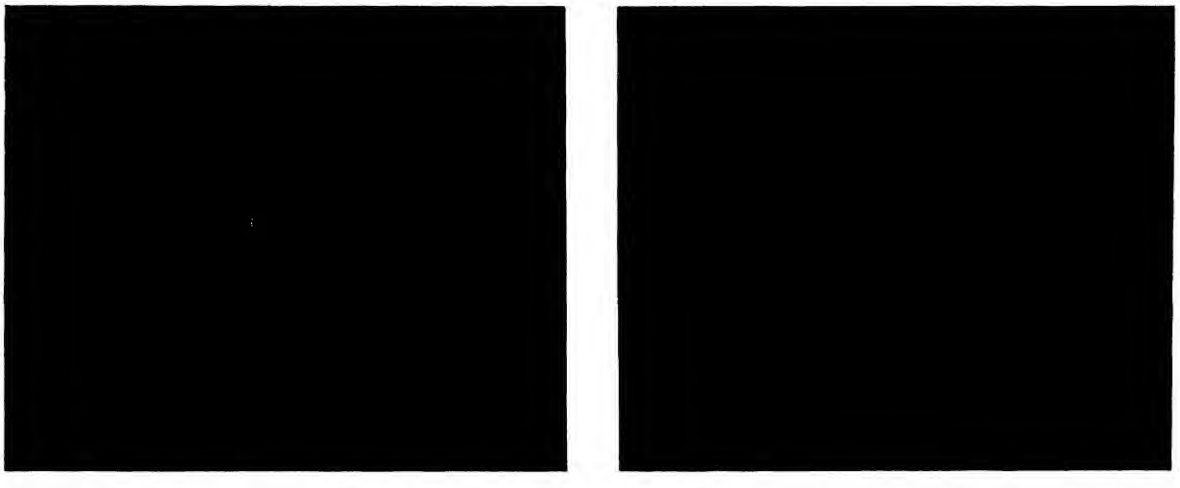


FIG. 2. Fiber patterns after relaxation at constant strain (*ca.* 26 percent). (a) Within 1 hour after termination of straining; (b) 48 hours later.

[3] S. D. Gehman and J. E. Field, J. App. Phys. 15, 371 (1944).
[4] W. L. Bragg, Proc. Camb. Phil. Soc. 45, 125 (1949).

in the fibers. If the orientation of the crystallites is broken down by heating or wetting of the fiber specimen, then the spots will diverge into more complete and uniform rings. As orientation or fibering takes place, the arcs converge into spots of smaller and smaller angular width. The interpretation of this broadness as a measure of the degree of orientation is clear and unambiguous.

Figure 3 shows a typical load-elongation curve for the Nylon used, when loaded at a constant time rate. After a small initial elastic rise followed by a somewhat rounded "yield point," the curve consists of two large portions; a concave upward region terminating in

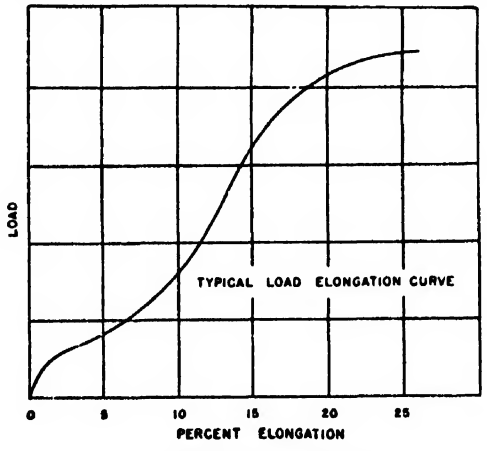


FIG. 3. Typical load elongation for Type 200 Nylon (constant r of load application). Load units arbitrary.

an inflection point at about 14 percent elongation, and a following portion concave to the strain axis which terminates at the rupture point.

Figure 4 is a plot of the results of the angular broadness measurement; they are plotted against the same strain ordinate as in the previous diagram. The broadness decreases up to strains at the inflection point of the load-elongation curve, at which there appears to be a cusp, followed by an increasing broadness to the rupture point. From the interpretation of angular broadness cited above, this plot suggests, for strains up to the inflection point of the load-elongation curve, that the crystallites rotate toward positions parallel to the fiber axis. Further, these data suggest that this process terminates at the inflection point and is replaced by a disordering process. This interpretation is supported by the data shown in Fig. 5. Here the radial broadness is plotted as a function of the elongation. For strains up to the inflection point of the load-elongation curve, the broadness remains approximately constant but rises rapidly for larger strains. From the shape of this curve and from Fig. 3, it seems rather difficult to attribute this radial broadening to tessellated elastic strains, existing throughout the irradiated volume, which would be expected to rise monotonically throughout the straining epoch. On the other hand, it appears logical to associate this with a reduction in particle size, that is, crystallite break-up. Analogous measurements of radial broadening associated with macroscopic plastic torsional strains have been made for metals by Patterson and Orowan.[5] In both the magnitude of the broadness change and in shape, their results are similar to those broadness values shown in Fig. 5 which are associated with strains greater than the inflection point of the Nylon load-strain curve, 14 percent. The view of these authors varies as to the details of interpretation of this result.

The intensity of the amorphous scattering was compared to the crystalline scattering and was found to decrease with strain up to the inflection point of the load-elongation curve, and to remain approximately constant thereafter. Since this measure was difficult to make accurately, we offer at present only the qualitative result.

DISCUSSION

The experimental result and its apparent interpretation which were obtained above can be used together with other known observations concerning plastic flowing in metallic solids to infer a very interesting and probably general characteristic of the reaction of reasonably well-ordered solids to large load applications.

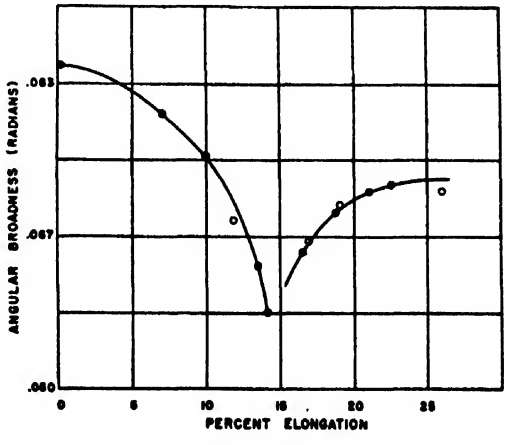


FIG. 4. Angular broadness *vs.* the elongation (%). Type 200 Nylon.

Recently, Koehler and Blewitt,[6] from studies of the changes in electrical resistivity with applied load, have inferred the existence of a sharp change in the micro deformation mechanisms at small plastic strains in ordered $AuCu_3$ single crystals. They have further pointed out that a similar inference can be obtained from: (1) a study of the curvature of "resolved" stress-strain relations of single crystals of $AuCu_3$, Cu, Ag, and Au, (2) x-ray line broadness changes with cold work in polycrystalline Cu, and (3) in other less direct ways. In view of the Nylon result reported above, the

[5] M. S. Patterson and E. Orowan, Nature 162, 991 (1948).
[6] J. S. Koehler and T. H. Blewitt, Phys. Rev. 75, 1952 (1949).

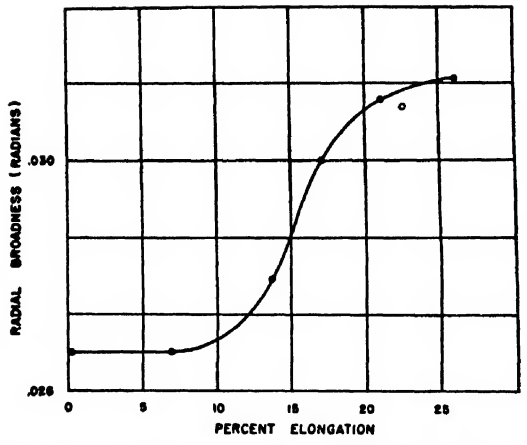


FIG. 5. Radial broadness *vs.* elongation (%). Type 200 Nylon.

Koehler-Blewitt generalization might be revised to the following. For any well-ordered solid, one will likely find in the region of true plastic flowing a narrow region of plastic strain through which there occurs a relatively sharp transition in the micro mechanisms of reaction to applied loads.

Recently, many theoretical papers have appeared which attempt to explain plastic flowing, and apparently related phenomena in metallic solids in terms of elementary units of crystallographic disorder termed "dislocations."[7,8] In view of the relatively high degree of order in the Nylon used for the work reported in this paper, one might seek an explanation of the results presented in terms of the dislocation concept. It should be noted, however, that the broadness measure reported here does not constitute evidence for the existence of particular dislocation arrays or, in fact, any array within a domain, but rather the existence of domains characterized by a linear dimension which is inversely proportional to the broadness. This dimension is the quantity which changes with the applied load. The domain viewpoint focuses, then, one's attention on the physically observed, whereas an explanation of the Nylon effects in terms of a generalized dislocation concept would stress entities which were not observed in the experiments. It is recognized that some type of disorder must exist between domains leading to x-ray line broadness, but it is felt that until experimental techniques are perfected which allow its study, explanations of general mechanical behavior in terms of disorder concepts should be avoided.

The suggestion that there is generally a transition in the microscopic mechanisms of plastic flowing in solids will, if verified, pose important theoretical problems. Currently, theoretical discussions of work hardening, e.g., in terms of dislocations often assume what is essentially an "exhaustion" hypothesis. Given amounts of the disorder are thought to initially exist. These theories explain the effects associated with continued plastic flowing in terms of the diminution of the number of elementary disordered regions with increased strain or time. Such exhaustion theories will surely require modifications if they are to account for the possibility of a transition in the micro reactions to gross flowing. The physical nature of flowing in solids, whatever it is, is surely extremely complex, even in what might be termed "simple" cases.

[7] G. I. Taylor, Proc. Roy. Soc. **A195**, 362 (1934).
[8] J. M. Burgers, Proc. Phys. Soc. **52**, 23 (1940).

CONCLUSIONS

The new data for highly oriented Nylon fibers presented indicate a change in the nature of plastic flowing, which occurs in the straining cycle at the point of inflection of the load-elongation curve (approximately 14 percent elongation). In the initial portion of the curve, the mechanism of flowing appears to involve rotation of the crystallites toward positions parallel to the fiber axis accompanied by an additional crystallization at the expense of the amorphous material initially present. The latter follows directly from the variations noted in the ratio of the intensity of the amorphous halo to that of the diffraction spots. This process terminates at the inflection point and is replaced by a mechanism of crystallite break-up, and disorientation.

The x-ray line broadness effects in Nylon, coupled with less direct effects observed in metallic solids which appear to be related to plastic straining all point to the existence in these solids of at least two types of microscopic reactions resulting from the load application. The experimental evidence, while not as direct as might be desired in some cases, seems to point to the existence of domains of nearly perfect atomic order, the motions and sizes of which are altered by the macroscopic flowing. Finally, from the domain viewpoint, it would appear that there are more general similarities between the micro reactions of well-ordered solids to applied loads than is commonly accepted. Obviously, domain structure on the atomic scale varies with materials, but it may be that this scale of structure is unimportant in describing at least some of the qualitative features of microscopic plastic flowing reactions.

ACKNOWLEDGMENTS

The authors are indebted to their colleagues G. R. Irwin and W. H. Sanders for many illuminating and helpful discussions. H. Friedman and L. S. Birks kindly made diffraction equipment in their laboratory available for this study.

Aluminum Oxide Replicas for Electron Microscopy Produced by a Two-Step Process

G. Hass and M. E. McFarland
Engineer Research and Development Laboratories, Fort Belvoir, Virginia
(Received October 26, 1949)

A two-step method is described for preparing aluminum oxide replicas of aluminum and non-aluminum specimens. In this method a thick layer of aluminum is evaporated onto the surface to be studied and pulled off the surface with Scotch Tape or cement covered glass. Aluminum oxide is formed on the side of the aluminum film which has been in contact with the specimen by anodizing in 3 percent ammonium tartrate at about 40 volts. The aluminum oxide film, which represents a positive replica of the specimen surface, is released from its substrate in dilute hydrochloric acid, washed with distilled water, and mounted for the electron microscope investigation. The applicability of the described method is demonstrated with electron micrographs of various specimens. The aluminum oxide films produced in ammonium tartrate are free of pores and grain structure. Films only 100 to 150A in thickness can be prepared and are suitable as temperature resistant supporting membranes for electron microscope and electron diffraction investigations.

ALUMINUM oxide coatings produced anodically in a suitable electrolyte have proved to be particularly useful as replica films for electron microscope studies.[1] Originally, however, the application of this method was restricted to aluminum and aluminum alloys on which anodic films can be formed directly. Only recently there was reported a modification of this method, a two-step process, applicable to studies of the surfaces of some non-aluminum materials.[2] In this process a piece of electrolytically polished pure aluminum is pressed against the surface to be studied; thus an aluminum replica is produced. On this aluminum replica a thin anodic film is produced and then removed chemically for examination in the electron microscope. Evidently this "engraving method" is only applicable to studies of extremely hard surfaces which are not damaged or changed by the high pressure required to press the aluminum against them.

By using a double replica method similar to the silver-collodion one first suggested by Zworykin and Ramberg,[3] a more general application of anodic aluminum oxide as replica film can be obtained. The specimen to be investigated is placed in a vacuum chamber and a thick film of aluminum ($t > 1\mu$) and a thin layer of magnesium ($t < 0.1\mu$) is evaporated onto its surface. The specimen is removed from the vacuum chamber and the magnesium-aluminum film is pulled off the surface with Scotch Tape or a cement covered glass slide. If the aluminum is sufficiently thick, the magnesium-aluminum layer can readily be removed from most specimens without damaging their surfaces. Aluminum oxide is now formed on the

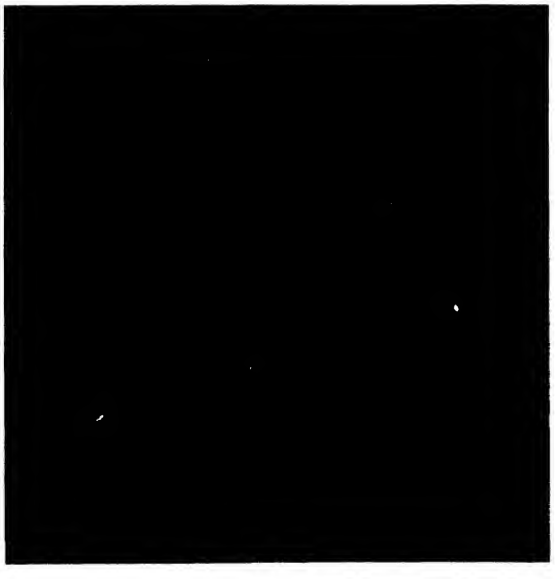


FIG. 1. Unshadowed aluminum oxide replica of a 30,000 lines per inch grating. 20,000×.

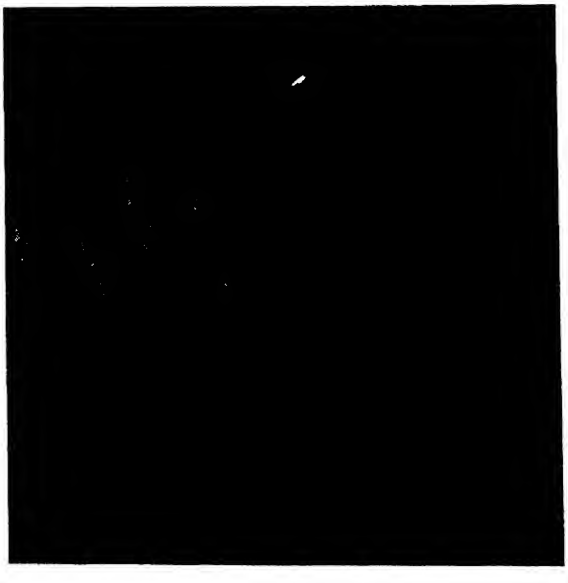


FIG. 2. Germanium-shadowed aluminum oxide replica of a 30,000 lines per inch grating. 10,000×.

[1] H. Mahl, Metallwirtschaft 19, 488 (1940) and Zeits. Tech. Phys. 22, 33 (1941).
[2] J. Hunger and R. Seeliger, Neue Phys. Blätter 2, 15 (1946) and Metallforsch 2, 65 (1947).
[3] V. K. Zworykin and E. G. Ramberg, J. App. Phys. 12, 692 (1941).

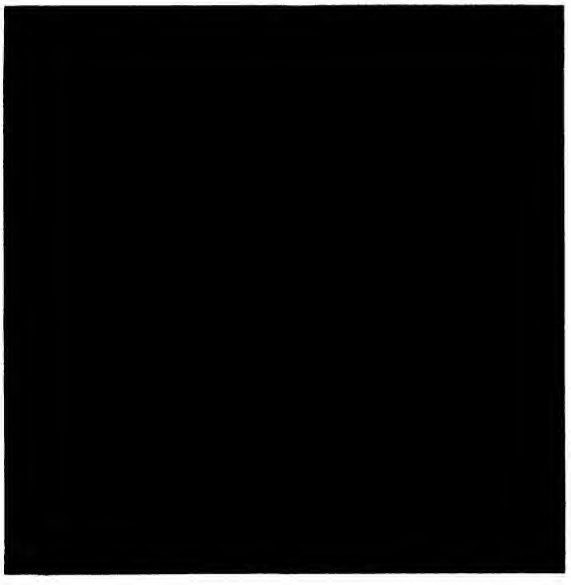


FIG. 3. Germanium shadowed aluminum oxide replica of a Plexiglass surface scratched by cleaning with cotton. 21,000×.

FIG. 4. Germanium shadowed aluminum oxide replica of a KRS-5 crystal on which KRS-5 has been condensed. 14,000×.

side of the aluminum film which has been in contact with the specimen by anodizing with 30–40 volts for 20–30 sec. The electrolyte used for the anodic oxidation is a solution of 3 percent tartaric acid with ammonium hydroxide added to make the *p*H about 5.5. The thin magnesium film placed between the aluminum and the Scotch Tape or cement makes the separation of the aluminum oxide replica easier and faster than it would be with aluminum only. If the film now consisting of aluminum oxide, aluminum, and magnesium on Scotch Tape is immersed in dilute hydrochloric acid the magnesium dissolves rapidly and the aluminum film with the oxide replica on top rises to the surface in less than a minute. There the aluminum also dissolves in a few minutes and the aluminum oxide replica film is left floating on the surface. After being washed in distilled water, the replica film can be mounted on a support screen for the electron microscope investigation. Replicas produced by the above method can be shadow cast to increase the contrast of surface details. Germanium has been used for shadowing because it evaporates easily and forms films with glass-like structure when condensed onto unheated bases.[4,5] Figures 1 and 2 are unshadowed and germanium-shadowed aluminum oxide replicas of a 30,000 lines/inch grating obtained by the aluminum-aluminum oxide method. The replicas were made from a reflection grating ruled on an evaporated aluminum mirror. It is remarkable that a thick film of aluminum evaporated onto the aluminum grating can be removed without damaging the grating. Aluminum-aluminum oxide replicas have been made from the original grating more than 30 times without changing its appearance. The unshadowed grating replicas (Fig. 1) show sharp dark lines and are therefore especially suitable for calibrating the magnification of electron microscopes and for measuring particle sizes exactly. The fact that anodic aluminum oxide films are of uniform thickness and of temperature resistant material makes the replicas withstand electron bombardment of high intensity. The germanium shadowed replica (Fig. 2) shows more contrast and detail and allows a better study of the grating surface. Figures 3 and 4 demonstrate the general applicability of the two-step replica method described above. Figure 3 shows an electron micrograph of a Plexiglass surface scratched by careful cleaning with cotton and Fig. 4 represents the surface of a KRS-5 crystal (TlBr-TlI) on which KRS-5 has been condensed slowly at high temperature.

The thickness of the oxide films formed anodically in a given time on aluminum in 3 percent ammonium tartrate increases linearly with the applied voltage. Layers produced in 30 sec. have a thickness of 12.2A-units per volt. After 30 sec. their thickness increases very little with the duration of the anodic treatment.[6] Such layers are amorphous and show no evidence of pores or grain structure. Films only 100 to 150A in thickness can be prepared. They are, therefore, suitable as temperature resistant support films for electron diffraction and electron microscope investigations.[7] Supporting membranes with extremely smooth surfaces can be prepared on electrolytically polished pure aluminum or on thick evaporated aluminum layers that have been stripped from a glass surface.

[4] H. König, Reichber. f. Physik 1, 4 (1944).
[5] G. Hass, Phys. Rev. 72, 174 (1947).
[6] G. Hass, J. Opt. Soc. Am. 39, 532 (1949).
[7] G. Hass and H. Kehler, Kolloid Zeits. 95, 26 (1941).

Diffusional Viscosity of a Polycrystalline Solid

CONYERS HERRING
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received December 23, 1949)

According to a suggestion of Nabarro, any crystal can change its shape by self-diffusion in such way as to yield to an applied shearing stress, and this can cause the macroscopic behavior of a polycrystalline solid to be like that of a viscous fluid. It is possible that this phenomenon is the predominant cause of creep at very high temperatures and very low stresses, though not under more usual conditions. The theory underlying it is developed quantitatively, and calculations of rate of creep, or equivalently of effective viscosity, are given for aggregates of quasi-spherical grains and for wires composed of cylindrical grains. Allowance is made for the effect of tangential stress relaxation at the grain boundaries. It is suggested that mosaic boundaries and boundaries between grains of nearly the same orientation may be unable to serve as sources or sinks of the diffusion currents, in which case the creep rate will depend only on the configuration of grain boundaries having a sizable orientation difference. Numerical comparison of the theory with experiments on the high temperature creep of wires favors this view, but is not entirely satisfactory. Suggestions for further experiments are made.

NABARRO[1] has recently pointed out that self-diffusion within the grains of a polycrystalline solid can cause the solid to yield to an applied shearing stress, the yielding being caused by a diffusional flow of matter within each crystal grain away from boundaries where there is a normal pressure and toward those where there is a normal tension, as shown in Fig. 1. This yielding is macroscopically describable by an effective viscosity proportional to the square of the linear dimensions of the grains. This phenomenon, which occurred independently to the present writer, provides a possible explanation for the quasi-viscous behavior which has been observed for foils and wires which are suspended with a very slight load and held at an elevated temperature.[2] It may also be important in other types of sintering phenomena. The purpose of the present note is to clarify certain points in the theory underlying this "diffusional viscosity" and to attempt a more refined quantitative calculation of its magnitude for several experimentally realizable situations.

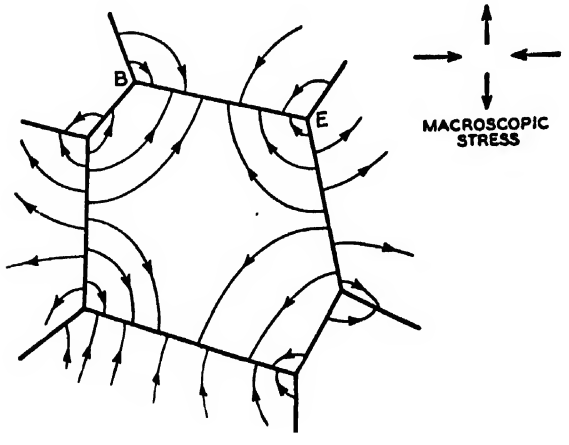


FIG. 1. A typical grain of a polycrystalline solid, showing the self-diffusion currents to be expected when the solid is subjected to a shearing stress.

1. BASIC THEORY

Let us consider first the simplest case of stress-motivated diffusion, namely self-diffusion in a single crystal of the cubic class containing only one type of atom and held at uniform temperature. In the absence of pressure gradients the diffusive flux of atoms, due to migration of lattice vacancies or interstitial atoms, will be proportional to the gradient of the concentration of these lattice defects; in the presence of pressure gradients, however, it tends to be energetically advantageous to move the lattice defects in whichever direction will relieve the inequality of pressure. It is easily shown that the effects both of concentration gradients and of pressure gradients are correctly given by writing the diffusive flux $\mathbf{j}$ (atoms per unit area per unit time) as proportional to the gradient of a chemical potential.[3,4] If n_l is the number of lattice sites per unit volume, D the self-diffusion coefficient, and T the absolute temperature,

$$\mathbf{j}=-(n_l D/kT)\nabla(\mu-\mu_h), \qquad (1)$$

where $(\mu-\mu_h)$ is the work per atom required for the reversible addition of a small quantity of material to the given region of the crystal, the material added being placed interstitially in the lattice or used to fill lattice

[1] F. R. N. Nabarro, "Report of a Conference on the Strength of Solids" (Physical Society, London, 1948), p. 75.

[2] *Foils:* J. C. Chapman and H. L. Porter, Proc. Roy. Soc. A83, 65 (1909); I. Sawai and M. Nishida, Zeits. f. Anorg. u. Allg. Chemie. 190, 375 (1930); G. Tammann and W. Boehme, Ann. d. Physik 12, 820 (1932). *Wires:* Udin, Shaler, and Wulff, J. Met. 1, 186 (1949); Alexander, Kuczynski, and Dawson, paper presented at the Symposium on the Physics of Powder Metallurgy organized by Sylvania Electric Products, Inc., New York, August 24–26, 1949.

[3] It has long been recognized that it is more appropriate to write the diffusion equations for multi-component systems with chemical potentials, or their equivalents, rather than with concentrations, and this fact has recently been the subject of a number of papers in the metallurgical literature. The most recent, and probably the most satisfactory, of these is that of J. Bardeen, Phys. Rev. 76, 1403 (1949).

[4] C. Herring, paper presented at the Symposium on the Physics of Powder Metallurgy organized by Sylvania Electric Products, Inc., New York, August 24–26, 1949. Plans have been made to publish the papers of this Symposium in book form.

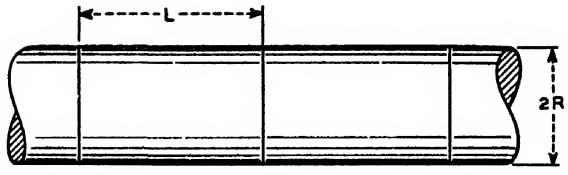


FIG. 2. Section of a wire with equally spaced transverse grain boundaries.

vacancies. This work is represented as a difference of two chemical potentials, μ and μ_h, to emphasize the fact that the change in question alters both the number of atoms and the number of lattice defects; however, in this paper only the difference of these quantities will appear.

Since any divergence of $\mathbf{j}$ within a single crystal would rapidly alter the concentration of lattice defects in such direction as to reduce $|\nabla\cdot\mathbf{j}|$, it will be safe to assume that in any continuing change $\nabla\cdot\mathbf{j}\propto\nabla^2(\mu-\mu_h)=0$. Thus the diffusion currents are completely determined if the values of $(\mu-\mu_h)$ on the boundary of the crystal are known. Now if the transfer of atoms from the interior to lattice sites at the boundary can occur easily, i.e., if the surface of the crystal is free to grow or recede with an accompanying change in the number of lattice defects in the interior, the boundary values of $(\mu-\mu_h)$ can be determined from the condition that the free energy of the crystal and its surroundings be a minimum with respect to transfer of atoms between any portion of the surface and the volume immediately beneath this portion. In reference 4 the boundary values of $(\mu-\mu_h)$ are determined in this way for the free surface of a crystal, and for grain boundaries where the interfacial tension between the grains is important; here we shall consider the simpler case where surface free energies can be neglected in comparison with the effects of externally applied forces. In intermediate cases the two effects have merely to be added.

Assuming that the surface-volume equilibrium just mentioned is realized at a grain boundary, the work involved in adding δN atoms to a crystal grain in the form of a hump on its surface must turn out to be the same, to the first order in δN, as the work involved in putting δN atoms into the interior just below the same surface region, holding the position of the surface fixed and placing the new atoms in interstitial positions or in places formerly occupied by lattice vacancies. The latter work is $(\mu-\mu_h-\mu_0)\delta N$, where μ_0 is the chemical potential of the source from which the δN atoms are obtained, say a piece of the same substance in equilibrium at the same temperature as the crystal being considered and at zero stress. The other process, formation of a surface hump, involves an amount of work $(\delta F-\mu_0\delta N-p_{zz}\delta v)$, where δF is the free energy of the hump itself, δv its volume, and p_{zz} is the normal traction on the surface, i.e., the zz component of the stress-tensor, where the z axis is chosen normal to the surface. Now to the first order in the stresses, $\delta F=\mu_0\delta N$, so equating the two amounts of work gives

$$\mu-\mu_h=\mu_0-p_{zz}\Omega_0, \tag{2}$$

where $\Omega_0=\delta v/\delta N$ is the atomic volume. This is the generalization of Nabarro's result regarding the effect of normal stress on vacancy concentration at the surface.

One can easily generalize Eqs. (1) and (2) to noncubic crystals or to crystals containing more than one kind of atom, provided the composition is uniform. In the multi-component case, however, care must be exercised in attempting to relate the D occurring in (1) to diffusion coefficients measured in other ways. Although a detailed discussion of diffusion in multi-component systems would be beyond the scope of this paper, an attempt has been made in the Appendix to indicate the type of analysis which must be used in applying the present equations to such things as substitutional solid solutions.

2. SPECIFIC CALCULATIONS

The rate of yielding of any given specimen to applied forces obviously depends on the detailed distribution of the sizes and shapes of its crystal grains, and on whether or not the grain boundaries are able to sustain shearing stresses for times as long as are required for appreciable diffusional transport to take place. It has been established in a number of cases that shearing stresses across metallic grain boundaries are rapidly relaxed at high temperatures;[5,6] it seems likely that this relaxation is a general property of grain boundaries, and that until this relaxation has become complete, the rate of creep due to it will usually be much faster than that due to diffusional viscosity. Although it is thus much more reasonable to compute the diffusional viscosity on the assumption of relaxed tangential stresses, we shall give calculations for both relaxed and unrelaxed tangential stresses, to show how much difference this factor makes. Both these calculations will be made assuming uniform, roughly spherical grains of a size small compared with that of the specimen; to show the way in which the rate of yielding is affected by having grains comparable with the dimensions of the specimen, some calculations will then be given for the rate of elongation of a wire consisting of roughly cylindrical grains end to end. Most of these cases are of course too idealized to apply accurately to any actual case, but it is hoped that they will sometimes be fair approximations to the truth.

Equal Quasi-Spherical Grains, No Grain Boundary Flow

If a crystal made up of grains such as that of Fig. 1 were suddenly subjected to a shearing stress, the initial diffusion currents would be very large in the neighborhoods of all sharp edges, such as E in the figure. This means that any grain boundary such as BE which is

[5] T. S. Kê, Phys. Rev. 71, 533 (1947); J. App. Phys. 19, 285 (1948); 20, 274 (1949).
[6] King, Cahn, and Chalmers, Nature 161, 682 (1948).

subjected to a normal tension will have more extra matter squeezed in at its edges than at its center. This will increase the tension at the center and decrease that at the edges, and a situation will very soon be reached where the normal flux is fairly uniform over each facet of the boundary. This suggests that for a crystal made up of nearly spherical grains all of the same size and having nearly isotropic elastic properties the rate of yielding can be fairly well approximated by that of a sphere acted on by such surface tractions as will produce a uniform shearing stress in its interior.

We shall assume for simplicity that the equations of Section 1 for a cubic crystal apply. Let the macroscopic stress-tensor be $\bar{p}_{\mu\nu}$. This must in all cases equal the volume average of the microscopic stress; in the present case we assume the stress in the interior of the sphere to be constant and equal to $\bar{p}_{\mu\nu}$. We assume $\bar{p}_{\mu\nu}$ to be a pure shear, so that

$$\sum_{\mu} \bar{p}_{\mu\mu}=0.$$

At a point of the surface of the sphere with coordinates x_μ the normal traction is

$$p_{rr}=\sum_{\mu,\nu} \bar{p}_{\mu\nu}x_\mu x_\nu/R^2,$$

where R is the radius of the spherical grain, so that

$$\sum_{\mu} x_\mu^2=R^2.$$

Inserting this into the boundary condition (2) it is easily seen that since $\nabla^2(\mu-\mu_h)=0$ the chemical potential in the interior must be given by

$$\mu-\mu_h=\mu_0-\Omega_0\sum_{\mu,\nu} \bar{p}_{\mu\nu}x_\mu x_\nu/R^2. \tag{3}$$

The normal flux at the boundary is, by (1)

$$j_n=(dr/dt)/\Omega_0=(2D/kT)\sum_{\mu,\nu} \bar{p}_{\mu\nu}x_\mu x_\nu/R^3 \tag{4}$$

since $n_l\Omega_0=1$. This may be compared with the rate of normal displacement of the boundary in a uniform shear of strain rate $\dot{u}_{\mu\nu}$:

$$dr/dt=\sum_{\mu,\nu} \dot{u}_{\mu\nu}x_\mu x_\nu/R. \tag{5}$$

Comparison of (4) and (5) shows that our assumptions are consistent with a uniform shear of all the grains of the crystal at a rate

$$\dot{u}_{\mu\nu}=(2D\Omega_0/kTR^2)\bar{p}_{\mu\nu}. \tag{6}$$

Viscous flow is defined, for the present case where $\bar{p}_{\mu\nu}$ is a pure shear, by the macroscopic relation $2\dot{u}_{\mu\nu}=\bar{p}_{\mu\nu}/\eta$, which with (6) gives for the effective viscosity of the polycrystalline solid

$$\eta=kTR^2/4D\Omega_0=(3/32\pi)^{\frac{1}{3}}(kT/D\Omega_0)V_g^{\frac{1}{3}}, \tag{7}$$

where V_g is the volume of a grain.

Equal Quasi-Spherical Grains, Tangential Stress Relaxed at Boundaries

For this case we must find a distribution of normal tractions over the surface of a sphere which will make the volume average of the stress-tensor in the interior equal the macroscopic stress-tensor $\bar{p}_{\mu\nu}$. Now from the equilibrium condition that the divergence of the stress-tensor must vanish we have, if $d\tau$ runs over the volume of the sphere and dS over its surface

$$0=\int x_\mu \sum_\alpha \partial p_{\alpha\nu}/\partial x_\alpha d\tau$$

$$=-\int p_{\mu\nu}d\tau+\int x_\mu \sum_\alpha p_{\alpha\nu}n_\alpha dS, \tag{8}$$

where $\mathbf{n}$ is the unit outward normal. For the present case

$$\sum_\alpha p_{\alpha\nu}n_\alpha=p_{nn}n_\nu,$$

so this gives

$$\bar{p}_{\mu\nu}=(3/4\pi R^4)\int x_\mu x_\nu p_{nn}dS. \tag{9}$$

If p_{nn} is represented by a series of spherical harmonics, (9) will determine the coefficients of the harmonics of order 2 (the lowest order consistent with the conditions of the problem) but will leave the higher order terms arbitrary. In an actual crystal grain these higher order terms will probably be no larger than is required to take account of the departure of the grain boundary configuration from spherical symmetry, and it is therefore proper to neglect them in the present approximation, the more so as they make no contribution to the over-all shear rate. Let us therefore set $p_{nn}=A(x_1^2-x_2^2)/R^2$, a form which can always be realized for the contribution of the second order spherical harmonics by proper choice of the coordinate directions x_1 and x_2. Insertion of this into (9) gives $\bar{p}_{11}=-\bar{p}_{22}=2A/5$, all other $\bar{p}_{\mu\nu}=0$. The chemical potential in the interior of the sphere is given by

$$\mu-\mu_h=\mu_0-(5/2)\bar{p}_{zz}\Omega_0(x^2-y^2)/R^2. \tag{10}$$

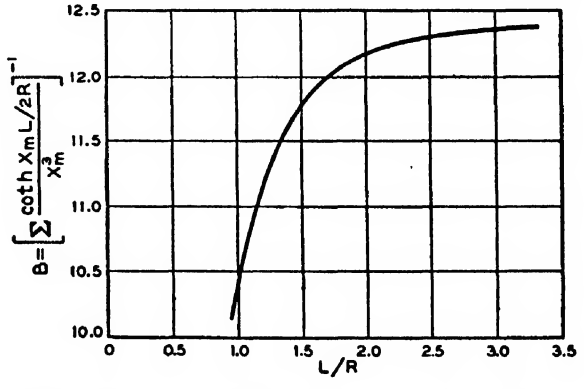


FIG. 3. Plot of the coefficient B occurring in Eq. (18) against the ratio of the length L of the grains to the radius R of the wire.

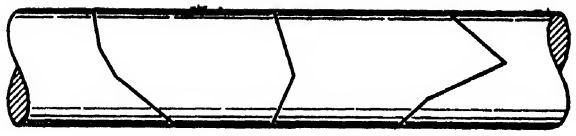


FIG. 4. Wire with irregular grain boundaries.

This is of the same form as (3) except for the factor 5/2. Repeating the argument leading to (7) gives

$$\eta = kTR^2/\tfrac{1}{10}D\Omega_0 = \tfrac{1}{10}(3/4\pi)^{\frac{1}{3}}(kT/D\Omega_0)V_a^{\frac{1}{3}}, \quad (11)$$

which is smaller than (7) by a factor $\frac{2}{5}$.

Wire with "Bamboo-Like" Grain Structure

Fine metal wires often consist of crystal grains occupying the entire cross section of the wire and having lengths which may be of the same order as the diameter of the wire or may be many times larger. In the experiments cited in reference 2 this type of grain structure characterized both the copper wires used by Udin, Shaler, and Wulff and the gold wires used by Alexander, Kuczynski, and Dawson.* A model which approximates the condition of these wires and for which the associated potential theory problem is easily solved is that of a wire with a "bamboo-like" grain structure, i.e., a cylinder of radius R with plane grain boundaries normal to the axis of the cylinder, uniformly spaced a distance L apart, as shown in Fig. 2.

For this model it is obvious from symmetry that when a longitudinal tension or compression is applied, the diffusion currents in the quasi-steady state will be such as to make the normal flux away from or into each grain boundary a constant over its area. This will of course require that the normal traction p_{nn} across a grain boundary vary with distance from the center of the wire; however, the integral condition

$$\int p_{nn}dS = F$$

must always be obeyed, where dS runs over the area of a grain boundary, and F is the force applied at the ends of the wire. For simplicity we ignore surface tension effects for the present; this is legitimate in computing the change in rate of stretching with a change in the load. In such case the boundary conditions on $(\mu-\mu_h)$ are:

$$f \equiv \mu-\mu_h-\mu_0 = 0 \quad \text{on the cylindrical surface} \quad (12)$$

while over the end boundary of each grain

$$\partial f/\partial z = \text{constant} \quad (13)$$

$$\int f dS = -F\Omega_0 \quad (14)$$

where z is the coordinate in the direction of the axis of the cylinder. The determination of f involves the same problem, therefore, as the determination of the potential due to an infinite series of uniformly charged disks placed inside a conducting cylinder. Using standard methods[7] the solution is easily shown to be

$$f = -\frac{F\Omega_0}{2\pi R^2}\frac{\sum_m [\coth(X_mL/2R)\cosh(X_mz/R)-\sinh(X_mz/R)]J_0(X_mr/R)/X_m^2J_1(X_m)}{\sum_m \coth(X_mL/2R)/X_m^3} \quad (15)$$

where the J's are Bessel functions, X_m runs over the roots of $J_0(X_m)=0$, and r is the distance from the axis of the cylinder, and where z ranges from 0 on one of the grain boundaries to $L/2$ midway between this boundary and the next. Each grain elongates or contracts at a rate

$$dL/dt = -(2\Omega_0/\pi R^2)\int j_z dS,$$

the integration being over the disk of radius R in the plane $z=0$. Using (1) and (15) we find

$$dL/dt = (D\Omega_0/2\pi R^3kT)BF, \quad (16)$$

where

$$B = [\sum_m \coth(X_mL/2R)/X_m^3]^{-1} \quad (17)$$

is a function of L/R plotted in Fig. 3. Its asymptotic value for $L \gg R$ is $B_\infty = 12.37$, and it does not begin to depart seriously from this value until L/R has become smaller than any value likely to be encountered frequently in wires whose crystal grains occupy the whole cross section.

In the present case the deformation of the wire cannot legitimately be described as viscous, because the grain size equals or exceeds the diameter of the specimen; however, it is interesting for comparative purposes to compute the viscosity η which a viscous filament of the same diameter would have to have in order to elongate at the same rate under the same load. For such a filament we have $F = \pi R^2 \cdot 3\eta d\ln L/dt$, and this agrees with (16) if

$$\eta = 2kTRL/3BD\Omega_0. \quad (18)$$

Wire with Irregular or Slanting Grain Boundaries, Tangential Stresses Relaxed

It is easy to show that if the grain boundaries are not normal to the axis of the wire as in Fig. 2, but are of the type shown in Fig. 4, the rate of yielding of the wire will

* I am indebted to the authors of these two papers for discussions of unpublished details of their work, including opportunities to examine photomicrographs of their wires, and for the statistics on mean grain lengths which have been used in Table I.

[7] See for example, W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939), Chapter 5.

be greater than that given by (16) for the same grain size, if the tangential stresses vanish at the grain boundaries. This is most easily seen by noting that a given dL/dt requires a given total normal flux integrated over the grain boundary,† regardless of its shape or inclination; in the electrostatic analogue this means a given total charge. The potential produced by this given charge, averaged over the grain boundary, will be numerically smaller when the grain boundary is oblique than when it runs normally across the wire as in Fig. 2. This potential is proportional to the normal stress across the grain boundary. It is easy to see that when the normal tractions are resolved along the axis of the wire, equal mean normal stresses will give equal values of the force F tending to elongate or compress the wire; therefore F will be less for Fig. 4 than for Fig. 2, when dL/dt is the same.

The calculations of the preceding section suggest some approximations which can be used to make a crude quantitative estimate of this effect. The same methods which were used to obtain (15) can be applied to calculate the potential due to a uniformly charged flat ring of any inner radius, r_1, and outer radius r_2, placed inside a grounded conducting cylinder. The potential problem for a grain boundary consisting of a series of ring-like steps, as shown in Fig. 5, can therefore be solved, and if the steps are not taken too large, this solution can probably be regarded as typical of irregular grain boundaries of the same average inclination. Ignoring interactions between the two ends of a grain, which Fig. 3 shows to be quite small, the force F required to produce a given elongation dL/dt at a single grain boundary of the sort described comes out to be

$$F=\frac{2\pi R^3kT}{D\Omega_0}\times\left[\sum_m\frac{\sum_{i,j}\{m,i\}\{m,j\}\exp(-|z_i-z_j|X_m/R)}{X_m^5[J_1(X_m)]^2}\right]\frac{dL}{dt}, \quad (19)$$

where i and j label the various rings, the ith ring having inner radius r_{i1} and outer radius r_{i2} and longitudinal coordinate z_i, and where

$$\{m,i\}=(X_mr_{i2}/R)J_1(X_mr_{i2}/R)-(X_mr_{i1}/R)J_1(X_mr_{i1}/R).$$

Now for large L the series in (17) differs from its leading term by only about 10 percent, so it is probably legitimate, in an approximation which replaces the grain boundaries of Fig. 4 by those of Fig. 5, to set the ratio of (19) to the F_0 given by (16) equal to the ratio of the first terms of the respective summations on m. This gives

$$F/F_0\approx\sum_{i,j}\{1,i\}\{1,j\}/X_1^2[J_1(X_1)]^2. \quad (20)$$

Some values computed from this formula for typical grain boundaries are given in Fig. 5.

The conclusion from this analysis is that the equivalent viscosity of a wire with irregular grain boundaries may be as much as two to three times smaller than the value given by (18) for flat-ended grains of the same length.

3. GRAIN BOUNDARIES *VERSUS* MOSAIC BOUNDARIES

The calculations of the preceding section have assumed the existence of certain "grains," at whose boundaries diffusion currents of lattice defects can converge or diverge, the boundary concentration of lattice defects being in equilibrium with the normal stress. It is conceivable that these properties might be possessed by crystal grains in the usual sense of the term, by mosaic blocks within a "single crystal," or, perhaps, by neither. In the present section it will be suggested as particularly likely that the condition of equilibrium just mentioned is attained, at sintering temperatures, on most ordinary grain boundaries, but is not attained on the boundaries of mosaic blocks. The experimental evidence discussed in the next section, though not entirely satisfactory, tends to favor this suggestion.

The boundaries of mosaic blocks have been pictured as consisting of rows of identical parallel dislocations spaced apart by a distance of the order of 10^3 atomic spacings.[8] Although it is not known whether the usual mosaic structure of single crystals consists of blocks bounded by such rows of dislocations, or merely of an irregular tangle of dislocation lines, it can hardly be doubted that the structure can be described in terms of dislocations; moreover, there seem to be good prospects for accounting quantitatively for many of the properties of grain boundaries, at least when the orientation difference between the grains is small, by picturing the boundaries as arrays of dislocations.[9] It is therefore

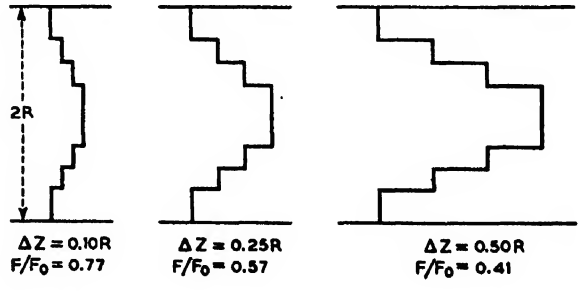


FIG. 5. Representation of irregular grain boundaries by successions of ring-like steps of height Δz at radii $0.3R$, $0.5R$, and $0.7R$. The quantity F/F_0 shown for each case is the approximate ratio of the force required to produce a given rate of elongation of the grain shown to the force required to produce the same rate of elongation of one of the flat-ended grains of Fig. 2.

† In accordance with the discussion given in Section 4, it is assumed that for sufficiently small stresses no "offsetting" of the grains will occur.

[8] J. M. Burgers, Proc. Phys. Soc. 52, 23 (1940); J. S. Koehler and F. Seitz, J. App. Mech. 14, A217 (1947).

[9] W. Shockley and W. T. Read, Phys. Rev. 75, 692 (1949); W. T. Read and W. Shockley, Phys. Rev. 78, 275 (1950).

legitimate to wonder what difference, if any, there is between the behavior of grain boundaries and that of mosaic boundaries.

In a well-annealed crystal the configurations of the dislocation arrays describing the mosaic structure must be such as to make the elastic free energy a minimum with respect to small displacements of any of the dislocation lines. Application of a small stress will affect the dislocations in two ways: a shearing stress on the slip plane of a dislocation will tend to move it in this plane, while a normal stress across a plane perpendicular to the slip vector will tend to move the dislocation in a direction not lying in the slip plane. Motion of the latter type requires the creation or annihilation of lattice defects, and will cease when the chemical potential of the lattice defects in the neighborhood of the dislocation has changed by an amount sufficient to compensate for the change in stress. An isolated dislocation line would thus affect the potential $(\mu-\mu_h)$ of the diffusion problem in much the same way as a wire held at a certain potential would affect an electrostatic problem. However, in an array of reasonably well-separated dislocations it seems likely that the freedom of motion of each dislocation will be impeded by the stress-fields of neighboring dislocations, and that within a short time after application of a small stress a new elastic equilibrium will be reached and all motion of the dislocations will cease, and concomitantly $(\mu-\mu_h)$ will become essentially constant over mosaic dimensions.

One aspect of this speculation can be given a rough check. The distance which each dislocation will move, in coming to a new position of equilibrium after application of a small stress σ, may be expected to be of the order of magnitude of $\sigma d/\sigma_0(d)$, where d is a length of the order of the spacing between dislocations and $\sigma_0(d)$ is the stress at a distance d from an isolated dislocation line. The resultant shearing strain of the whole specimen should therefore be of the order of this quantity times a/d^2, where a is the lattice constant. Since $\sigma_0(d)$ is of the order of $Ga/2\pi(1-p)d$, where G is a shear constant and p a Poisson's ratio,[10] the resulting shear is of order $2\pi(1-p)\sigma/G$. This gives values two or three orders of magnitude smaller than the total strains observed in the wire experiments of reference 2, which were of the order of 10^{-2}.

It is therefore not implausible to suppose that motion of dislocations may be of negligible importance in these experiments, and that the diffusion currents are divergenceless except at grain boundaries separating two crystals having a sizable difference of orientation $\Delta\theta$. By "sizable" is meant an orientation difference large enough so that the grain boundary can no longer be viewed as an array of dislocations spaced many lattice constants apart. If this speculation should turn out to be true, it would imply the existence of a critical value of $\Delta\theta$ below which growth of the two crystals normal to the grain boundary cannot take place freely, but above which it can. In such case any grain boundaries separating grains with less than the minimum $\Delta\theta$ should not be counted as grain boundaries in comparing the present theory with observations. There may conceivably be many such boundaries in wires whose grains have a strong tendency to crystallize with a certain crystallographic direction along the axis of the wire. For two grains with such a common axis need only a relative rotation about the wire axis to bring their orientations into coincidence, and it is possible that at sintering temperatures the natural desire of the grain boundary to decrease its interfacial tension may bring about such a rotation, at least until the orientation difference becomes small enough so that further changes of configuration cannot take place freely.

[10] See for example, J. S. Koehler, Phys. Rev. **60**, 397 (1941).

4. RELATION OF THEORY TO EXPERIMENT

Most of the present section will be devoted to a discussion of experiments on the longitudinal creep of wires at high temperatures, as measured by Udin, Shaler, and Wulff[2] for copper and by Alexander, Kuczynski, and Dawson[2] for gold. In these experiments an attempt was made to find what value of the applied tension would just suffice to prevent the wires from shrinking under the action of their surface tension, and thus to measure the latter quantity. The stresses used in these experiments were therefore below the range used in most ordinary creep experiments (about 1 to 100 g/mm^2), while the extensions were relatively large, of the order of 10^{-3} to 10^{-2} and more. It is therefore not unreasonable to attribute the creep in these high temperature experiments to mechanisms quite different from those operative in most of the previous creep experiments in the literature, although it is of course quite conceivable that the mechanisms may be the same, the thresholds being lower and the rates faster at the higher temperatures. The latter possibility is supported by the fact that Miller[11] has observed an apparently unlimited creep for single crystals of zinc under stresses of only 2 or 3 g/mm^2, even at room temperature; however, it would be risky to conclude from this that the experiments of reference 2 involve the same mechanisms as room-temperature creep in single crystals, and the following discussion will explore the alternative that the diffusion mechanism discussed in the previous sections plays a major role. As Nabarro has shown,[1] the latter mechanism does not seem to be the major one in creep experiments such as those of Chalmers[12] on single crystals of tin.

Offsetting and Kinking

Besides the diffusion effect and the possible effect of ordinary creep within the individual crystal grains, there are at least two other effects which may influence the changes in length which are observed for fine wires

[11] R. F. Miller, Trans. A.I.M.E. **122**, 176 (1936).
[12] B. Chalmers, Proc. Roy. Soc. **A156**, 427 (1936).

suspended at high temperatures. One of these is "offsetting" of adjacent grains, as shown in Fig. 6. This is known to occur for polycrystalline wires suspended at high temperatures under tension of sufficient magnitude.[13] However, it is easy to see that it cannot occur if the tension is below a certain limit, and that it cannot contribute to the shrinking of wires suspended with no load. For, referring to Fig. 6, consider a lateral offset of amount AB along a grain boundary whose normal makes an angle θ with the axis of the wire. This enables the wire to yield to the applied tensile force F by a distance $\langle AB\rangle \sin\theta$; however, it exposes an area of new surface represented by AB and CD. The surface free energy thus increases by $2R\cdot\langle AB\rangle\cdot(2\bar{\gamma}-\gamma_b)$ where $\bar{\gamma}$ is an average surface tension for the regions AB and CD, and γ_b is the specific free energy of the grain boundary. Thus the over-all free energy will be lowered by the offset if and only if

$$F>2R(2\bar{\gamma}-\gamma_b)\csc\theta. \tag{21}$$

This is somewhat larger than the force $\pi R\bar{\gamma}'$ necessary to prevent shortening of a long grain under its own surface tension. We may therefore conclude that offsetting will not occur in the lower range of stresses employed in the wire experiments of reference 2, but that it may well occur in the higher part of the stress range.

The other possible perturbing effect, which will be called "kinking," consists in the development of a slight angle between the axes of adjacent grains of the wire. This could be produced by a relative rotation about an axis normal to the wire axis. Such a rotation might be motivated by a dependence of the free energy of the grain boundary on the relative orientations of the grains; it would of course require that material be removed from the portions of the grains which move toward each other, and added to the portions which move apart. If this transport of matter takes place by diffusion, one can estimate the order of magnitude of the rate of change of orientation by the methods described in reference 4. This estimate, the details of which will not be given here, shows that for the wires of reference 2 the rate of kinking due to this cause ought to be too small to produce a perceptible shortening, though possibly not too small to give a measurable angle of kink.

Kinking might conceivably also be caused by a relative rotation of two grains about the normal to a plane grain boundary, if this normal were a trifle inclined to the axis of the wire. The motivation could again come from an anisotropy of the grain boundary energy; for this process, however, no transport of matter would be required, and its rate might be faster than that of the process described in the preceding paragraph.

Calculated *versus* Observed Creep Rates

Table I gives a comparison of the observed rates of creep of the wires of reference 2, expressed in terms of equivalent viscosities, with rates calculated from (18) and (20). Since the least accurately known of the experimental data are the lengths L of the crystal grains, the comparisons are made on this quantity.

Ideally, if the diffusion hypothesis is the correct one, it should be possible to get a fairly accurate check between theory and experiment. For, with one possible exception, all the quantities on the right of (18) can be measured fairly accurately, the most uncertain being D, and even this can be measured to an accuracy of 20 percent or better if the temperature can be accurately reproduced. The possible exception is the mean length L of the grains: if, following the possibility suggested in Section 3, it turns out that growth can occur at grain boundaries where the orientation difference is large but not at those where it is small, then the value of L to be inserted into (18) will be larger than the value L' measured metallographically, by an amount which it may be hard to establish. The correction factor F/F_0 will in many cases be close to unity, since both theory and observation agree that grain boundaries usually prefer to align themselves normal to the wire axis; if for any reason the boundaries are not transverse, a reasonable estimate of F/F_0 can be made from Fig. 5.

A number of experimental uncertainties in the present examples make the accuracy of the comparison of theory and observation fall somewhat short of this ideal. Besides the uncertain possibility that L may be appreciably greater than L', the value of L' itself is somewhat uncertain. The values given in the table were taken from photomicrographs of short sections of the wires after etching; however, since regions of long grains may alternate with regions of short grains, many such samples must be measured to get a reliable measure of the mean length of all the grains in the specimen. Moreover, the grain structure may change appreciably during the experiment. In this connection it is interesting to speculate that a relative axial rotation of neighboring grains, resulting ultimately in making the orientation difference too small for growth to occur, might conceivably cause the effective L to increase with time. This would lead to a creep rate decreasing with time, as was observed by Alexander, Kuczynski, and Dawson.[2]

A moderate error in the temperature, or rather, in the relative temperature scales of the creep and diffusion measurements, can produce an appreciable error in D. Dr. Alexander has informed the author that the temperature listed in the table as 1050° may have been closer to 1025°; a temperature lowering of this amount would

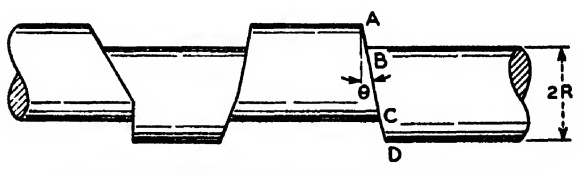


FIG. 6. Illustration of offsetting for a wire suspended under tension.

[13] See for example, C. J. Smithells, *Tungsten* (D. Van Nostrand Company, Inc., New York, 1936), second edition, p. 81.

TABLE I. Comparison of observed viscosities of wires with predictions of Eqs. (18) and (20).*

Authors†	USW		AKD
Metal	Cu	Cu	Au
T°C	1000	1000	1050
R, microns	36	64	14
D, cm²/sec.‡	2.5×10^{-9}	2.5×10^{-9}	8×10^{-9}
Observed η, poises	4.4×10^{12}	3.2×10^{12}	1.4×10^{12} (initial) 2.4×10^{12} (final)
$(L/R)(F/F_0)$, calculated	1.0	0.24	9.6 to 16
L'/R, observed	4	2	2

* The observed η-values for copper were taken from lines drawn by Udin, Shaler, and Wulff averaging their measurements at different temperatures. Their values must be multiplied by 2 because of an error in a formula, derived by Frenkel, by which they reduced their observations.

L is the average spacing between grain boundaries at which growth can occur, while L' is the average spacing between observable boundaries.

F/F_0 is the correction factor to be applied to (18) due to departure of the grain boundary shapes from the ideal transverse planes of Fig. 2; according to (20) this factor is probably between $\frac{1}{2}$ and 1.

The next to the last line of the table was computed by equating the observed viscosity η to the product of F/F_0 by the right of (18) taking B to be 12.2, and solving for $(L/R)(F/F_0)$.

† See reference 2.

‡ The self-diffusion coefficient tabulated for copper is close to the values measured by Steigman, Shockley, and Nix, Phys. Rev. **56**, 13 (1939); M. S. Maier and H. R. Nelson, Trans. A.I.M.E. **147**, 39 (1942); and Raynor, Thomassen, and Rouse, Trans. Am. Soc. Metals **30**, 313 (1942). That for gold was obtained by extrapolating the data of H. A. C. McKay, Trans. Faraday Soc. **34**, 845 (1938).

lower D, and therefore the calculated $(L/R)(F/F_0)$, by a factor of about 1.5.

If the diffusion of a radioactive tracer isotope occurs largely or predominantly by simultaneous interchange of positions by neighboring atoms, the D to be used in the present calculations will of course be less than that measured in the tracer experiment, perhaps very much less.

Remarks on Future Experiments

The comparisons given in Table I, though not very satisfactory, do at least show that the creep observed in the experiments of reference 2 is of the same order of magnitude as that predicted on the hypothesis that it is due to diffusion currents flowing between the free surface and the grain boundaries and that it is therefore distinctly slower than that which would be predicted by an analogous theory in which the diffusion currents converge on the boundaries of mosaic blocks with dimensions of the order of 10^{-4} cm. However, an obvious direct check of the hypothesis that grain boundaries are essential could be made by performing the same experiments with single crystal wires, or with wires having a very large L/R ratio. In such an experiment it would be desirable to check the orientations of different parts of the "single crystals" by x-ray methods, to detect the possible presence of a polycrystalline structure with grain orientations differing by only a few degrees.[14]

Several items may be mentioned which would facilitate the interpretation of future experiments with the usual polycrystalline wires. Grain lengths should be measured over the whole length of the wire, and if possible the distribution of orientation differences between adjacent grains should be checked. Checks should also be made to find out whether the sizes and orientations of the grains change appreciably during the experiment, and whether changes in the creep rate with time are permanent or recoverable. Offsetting and kinking should of course be watched for, though according to the discussion above they are not expected to play a major role in the stress range of most interest.

Diffusional viscosity should not be very greatly affected by small amounts of impurities, once the effect of these impurities on the grain structure has been allowed for. This could easily be checked.

An interesting consequence of the theory of this paper is that the cylindrical surfaces of the wires used in experiments such as those of reference 2 ought to develop shallow depressions along the grain boundaries when they lengthen with time, and ridges when they shrink. This is because the diffusion currents tend to be especially strong at the parts of the surface near the grain boundaries. Use of polished wires would facilitate detection of this effect.

In conclusion, it should be pointed out that the phenomenon of diffusional viscosity, if it turns out to be the predominant cause of creep under very low stresses at sintering temperatures, should be very important in all kinds of sintering processes. This is especially true with regard to the final stages of sintering: the shrinking down of the cavities requires an overall contraction of the specimen, and this can easily be achieved if the crystal grains are free to slide on each other tangentially and to grow or recede normally to their boundaries. In the absence of such processes and of plastic flow within the crystals, the only mechanism which could shrink the cavities would be a diffusion of material from the outer boundary of the specimen, as postulated in the theory of Pines.[15] The latter process obviously gives a rate of densification much slower than the former process, and one which decreases rapidly with increasing dimensions of the specimen.

APPENDIX

In a cubic crystal containing various kinds of atoms, the general expression for the diffusive flux of the ith kind of atom is

$$\mathbf{j}_i = -(n_L/kT)\sum_k D_{ik}\nabla(\mu_k-\mu_h), \qquad (A)$$

where n_L is the number of unit cells in unit volume. Immediately after stresses are applied, the currents of the different species will in general be widely different; however, before any macroscopically recognizable transport of material has occurred concentration gradients will be set up in such directions as to tend to equalize these differences in currents, and if the stresses are small a quasi-steady state will very soon be reached in which $\mathbf{j}_i = X_i\mathbf{j}$, where X_i is the fractional concentration of the ith species and $\mathbf{j} = \sum \mathbf{j}_i$. It will be shown that $\mathbf{j}$ can be expressed in terms of the gradient of a single harmonic function. Equating (A) to $X_i\mathbf{j}$ and solving for $\nabla(\mu_k-\mu_h)$,

$$\nabla(\mu_k-\mu_h) = -(kT/n_L)\mathbf{j}\sum_i (D^{-1})_{ki}X_i, \qquad (B)$$

[14] Such structures have been observed in a "single crystal" of aluminum by P. Lacombe, "Report of a Conference on Strength of Solids" (The Physical Society, London, 1948), p. 91.

[15] B. Y. Pines, J. Tech. Phys., U.S.S.R. **16**, 737 (1946). I am indebted to Dr. G. C. Kuczynski for a translation of this paper.

where D^{-1} is the reciprocal of the matrix D. Now define μ_{0k} at each point of the crystal as the value of μ_k corresponding to the composition occurring at that point but at zero stress. Then by the Gibbs-Duhem relation, $\sum X_k \nabla \mu_{0k} = 0$. Moreover, if the stresses are small ($\mu_k - \mu_k - \mu_{0k}$) will be a small quantity of the first order in the stresses. Multiplying (B) by X_k and summing gives

$$-(kT/n_L)\mathbf{j} \sum_{i,k} (D^{-1})_{ki} X_i X_k = \sum_k X_k \nabla(\mu_k - \mu_k - \mu_{0k})$$

$$\approx \nabla \sum_k X_k(\mu_k - \mu_k - \mu_{0k}) \quad \text{(C)}$$

if we neglect quantities of the second order in the stresses, since ∇X_k is a small quantity of the first order. Thus we can write, in formal analogy to (1),

$$\mathbf{j} = -(n_L D'/kT)\nabla f \quad \text{(D)}$$

with

$$f = \sum_k X_k(\mu_k - \mu_k - \mu_{0k}) \quad \text{(E)}$$

and

$$D' = [\sum_{i,k} (D^{-1})_{ik} X_i X_k]^{-1}. \quad \text{(F)}$$

An argument of the same type as that leading to (2) shows that at the boundary of the crystal

$$f = -p_{as} \sum_k X_k(\partial v/\partial N_k) = -p_{as}\Omega_0, \quad \text{(G)}$$

where $\partial v/\partial N_k$ is the partial atomic volume of the kth component and Ω_0 is the mean atomic volume.

The form of the matrix D_{ik} of course depends on the mechanism of diffusion, a subject beyond the scope of this paper. We may merely mention that for purely interstitial diffusion the matrix is diagonal, so that for a two-component crystal

$$D' = (X_1^2/D_{11} + X_2^2/D_{22})^{-1},$$

while for a two-component crystal in which there is no diffusion except by place interchange the determinant $|D_{ik}|$ vanishes, making $D' = 0$, as is to be expected for this model.

Streak Photography

IRWIN VIGNESS AND R. C. NOWAK
Mechanics Division, Shock and Vibration Branch, Naval Research Laboratory, Washington, D. C.
(Received December 5, 1949)

A description is given of some methods of "streak photography" used for obtaining displacement records as functions of time. Ordinary streak photography consists of photographing a moving object on a moving film camera with an open shutter. The motion of the object should be in a plane parallel to the film and perpendicular to its motion. Large errors in velocity determinations may occur if the motion of the object has components parallel to the film motion. A modified method of streak photography is described in which these errors are reduced to small values. By this method time-displacement curves have been recorded for displacements of many inches and periods of several seconds with accuracies within a few thousandths of an inch and a few hundred thousandths of a second.

INTRODUCTION

THE measurement of transient motions of mechanical elements is of constant necessity in laboratory and field problems associated with mechanical shock and vibration. The conditions of the experiments require methods of measurements that are easy to perform and that do not appreciably load the mechanical parts under study. While high speed moving pictures provide this together with a wealth of other information, the work of transcribing the motion picture records into quantitative displacement curves is too laborious for many conditions. Methods of streak photography are easy to perform and can often be done with ordinary laboratory equipment. Many types of oscillographic recording cameras can be used with little or no modification. Several Naval Laboratories have made extensive use of the methods of streak photography to conveniently obtain accurate displacement curves.[1] However, no descriptions have been made of these methods except as incidental to other experiments. This paper describes the methods of streak photography most generally used.

[1] Vigness, Kammer, and Holt, "Shock and vibration instrumentation and measurements," Second Partial Report, Naval Research Laboratory Report No. 0-2645 (September, 1945).

ORDINARY STREAK PHOTOGRAPHY

One of the simplest methods of recording transient motions of an object is to photograph the object with a moving film camera. The film of the camera is made to move at a known speed and the shutter is held open during the entire exposure. The camera is arranged so that the direction of motion of the object under study is perpendicular to the camera optic axis and to the direction of film motion. Sufficiently good contrast can often be obtained without any special preparation of the object, but usually it is advisable to arrange bright spots with a dark background on the object whose motion is being studied. These bright spots may be dots of white paint, ball bearings, or small luminous sources. Ball bearings are particularly advantageous if very fine lines are desired. A small light source some distance from the ball should be used, the reflection of this light by the ball closely approximates a point. The ball should be small so that the reflected light will continue to come from essentially the same location on the ball. Many spots may be in the field of view, and, consistent with the requirements of depth of focus, they can be spread over any distance from the camera. Of course the displacement scale will depend upon this distance. Illumination

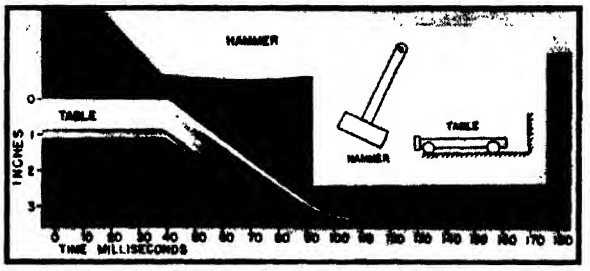


FIG. 1. Ordinary streak photograph showing the horizontal components of a displacement of a hammer and table when the table is struck by the hammer.

can be conveniently obtained for the non-luminous spots by photo-flood lights of the R3 type.

An example shown in Fig. 1 illustrates a hammer striking a steel table. The hammer swung as a pendulum and struck the edge of the steel table which was mounted on rollers and was constrained to move in but one direction. The upper part of the figure shows the motion of the hammer. The impact occurs at the sudden change of slope. The velocity of the hammer and table before and after impact can be determined from the slopes of the corresponding streaks. The boundary streaks of the hammer and table do not come into contact at impact as the camera did not observe the immediate areas of impact.

Special methods of streak photography have been applied for measuring the rate of flame propagation and for determining the velocity of the luminous areas of explosions. A precise and very accurate method of determining the deceleration of a projectile, and hence the force exerted by the projectile, is routinely performed by the Ballistics Branch of the Naval Research Laboratory by an unusual streak photographic technique.[2] In this case the moving film is contained on the inside of a drum that is 4 inches in diameter and that rotates so that film speeds up to 600 feet per second are attained. An intense source of light is flashed on for a half-millisecond during the passage of the projectile and the shadow of one point of the base of the projectile is recorded on the moving film.

The method of streak photography results in a direct plot, in Cartesian coordinates, of the displacement time history of as many locations as can be contained in the field of view. It possesses the advantage over high speed moving pictures of greater accuracy and the elimination of point by point plotting of curves when displacement records are required. High speed moving pictures, of course, possess many advantages in other respects.

Accurate time scales can be obtained by a variety of methods. Perhaps the most obvious is to place a stationary arc in the camera field of view. "Point" arc light sources similar to that supplied by Central Scientific Company can be controlled by an amplified electrical signal that has been generated by a tuning fork. When a camera slit is used, as for the modified method of streak photography described below, it is convenient to reflect the light from a "Strobolux" (General Radio Company) into the slit so that timing lines are formed at regular intervals. Some oscillographic cameras use spark discharges that expose the film adjacent to the spark. Others have an intermittent light source internal to the camera and focused on the film. Rotating drum cameras may have their film speeds determined simply by measuring the rotational speed of the drum. Generally, little difficulty is encountered in obtaining a precise time scale even if the film speed is non-uniform.

MODIFIED STREAK PHOTOGRAPHY

A source of error in the ordinary method of streak photography lies in the possibility of small motions of the object in a direction parallel to the film motion. In order to eliminate this error, to measure large displacements (4 or 5 inches), and to obtain an accuracy within a few thousandths of an inch, a modified method of streak photography was devised. The principles of this method are shown in Fig. 2. In this figure the object consists of a series of bright lines on a dark background. The lines are contained in a plane parallel to the direction of the film motion. The direction of motion of the object, which is rigidly attached to the part whose motion is being studied, should be perpendicular to the camera optic axis and to the direction of film motion. The camera lens will focus an image of these lines on the film. A slit and a cylindrical lens (the lens is optional) placed near the film focuses the line to an approximate point. Motions of the object in a direction parallel to the film motion will not result in any displacement of the image on the film. Small rotations in a plane perpendicular to the camera axis will produce second-order errors, and motions out of this plane will result in the usual scale changes and depth of focus problems. Such an arrangement, used at the Naval Research Laboratory, had an object that consisted of a grid of music wire; the wires of which were about 0.01 of an inch in diameter and were spaced 0.200

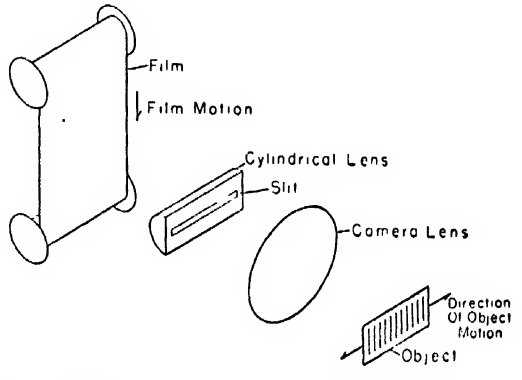
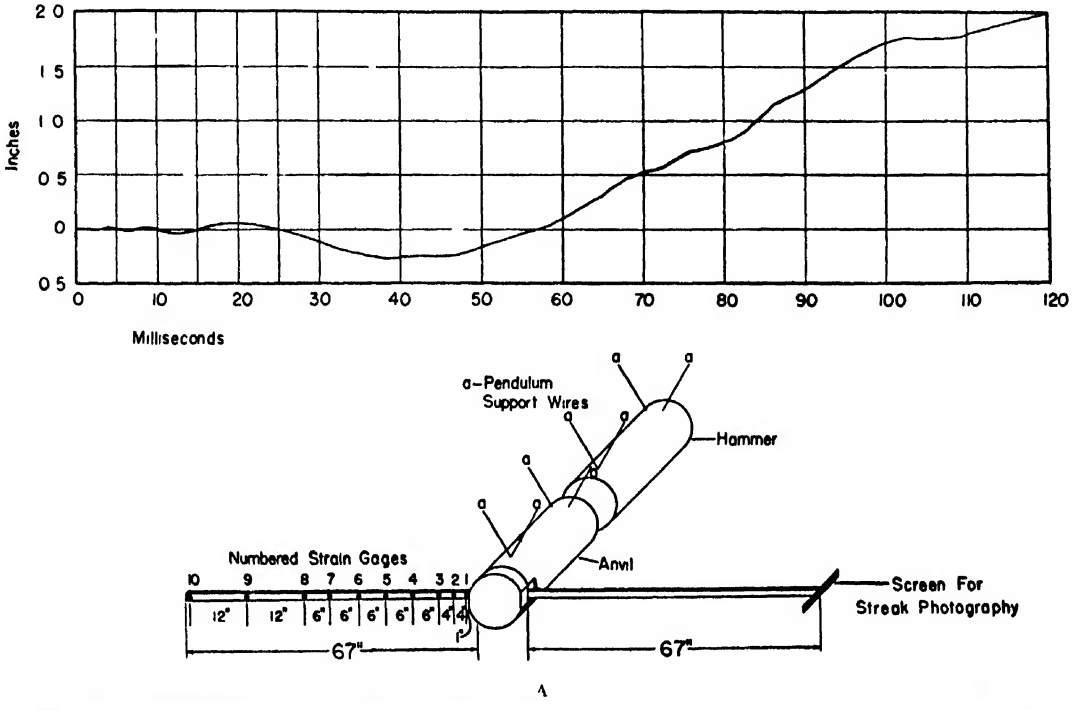


FIG. 2. Principal parts used for the modified method of streak photography. The slit should be near to the film compared with the film-lens distance.

[2] A. Victor Masket, "The measurement of forces resisting armor penetration," J. App. Phys. **20**, 132–140 (February, 1949).

of an inch apart in narrow grooves that were machined in a black Bakelite base. The object was varnished so as to cement the wire in position. The use of a wire for the modified method of streak photography is analogous to the use of a ball for the ordinary method and makes possible a very fine line image. Reasonably good results were also obtained by simply filling sharp machined scratches with white ink in a similar Bakelite base. A convenient camera arrangement consists of a 2-inch focal length lens attached to a moving film oscillograph camera (General Radio GR 651A-E). An extension tube was usually used that held the lens about 9 inches from the film. This arrangement caused the object to be enlarged about 4 times in linear dimensions and the image of 2 object wires could be seen on the 35-mm film when one image was near the edge of the film. This fixed the displacement scale on the film. A cylindrical lens was cemented over a $\frac{1}{16}$-inch wide slit which replaced the



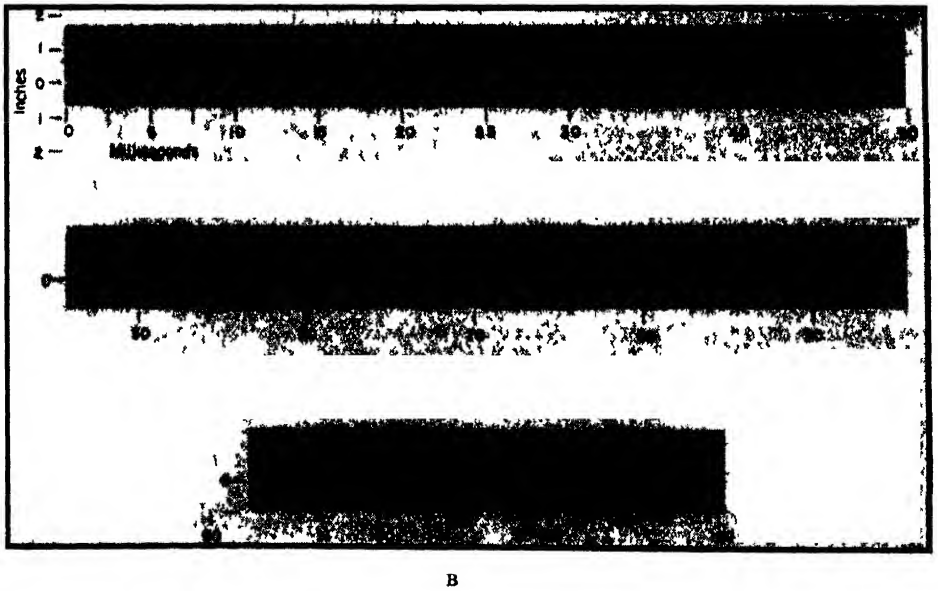


FIG 3. Displacement curves obtained by streak photography. The upper curve was derived from the photographic record shown on part B of the figure.

ordinary film gate. Fair pictures may be had with a $\frac{1}{32}$-inch slit and no cylindrical lens. When wires were used for the grid lines the best results were obtained with a small source of illumination, such as a laboratory arc with a focusing lens. The light must of course be in a plane perpendicular to the grid lines and containing the lens optic axis. Sufficient illumination is easily obtained for writing speeds up to 11 ft./sec.

Figure 3 illustrates this method as used to determine the motion of the free end of a cantilever beam when the built-in end suddenly acquires and maintains a uniform velocity. The experimental arrangement is shown in the center part of this figure. The hammer strikes the anvil

FIG. 4 An instrument shock testing machine with attached grid for streak photography.

and imparts a velocity to the anvil that remains essentially constant for the duration of the experiment. The velocity is in the direction of the anvil axis and the displacement is measured in this direction. A reproduction of the streak photograph is shown on the lower part of Fig. 3. The displacement scale is determined by the vertical separation of lines which represents 0.2 inch on the object. A continuous displacement curve, such as is shown on the upper part of this figure results from

FIG 5 Streak photograph illustrating the motion of the platform of the instrument shock testing machine.

joining together the various curves, shown on the lower section, with proper multiples of the 0.2-inch vertical separation.

Figures 4 and 5 show the application of the modified method of streak photography for the determination of the motions of a shock testing mechanism. As shown on Fig. 4 a small platform falls on a stiff spring and rebounds. The vertical motion of this platform is shown through the impact time in Fig. 5. The object grid in this case consisted of white inked grooves that were machined at 0.2-inch intervals.

SUMMARY AND CONCLUSIONS

Ordinary streak photography which consists of photographing a moving object on a moving film shutterless camera is a convenient and accurate method of obtaining displacement curves as a function of time provided the direction of motion studied is previously known and is unidirectional. When there are components of motion, which result in image displacements parallel to the motion of the recording film, a modified method of streak photography must be used. In this method a small rigid body is attached to the object whose motion is being studied. One or more parallel lines, which are made perpendicular to the principal component of motion and which are contained in the focal plane of the camera, are scribed on this body. The image of these lines on the camera film is cut to an approximate point by introducing a slit near the film or by introducing a slit and cylindrical lens a short distance from the film. By the above method transient displacements of many inches can be measured to an accuracy within a few thousandths of an inch. The time scale can easily be extended to more than a second with an accuracy up to about 10 microseconds.

Surface Tension Effects in Oriented Thin Silver Films*

G. W. JOHNSON†
Department of Physics, State College of Washington, Pullman, Washington
(Received November 7, 1949)

Structural changes of thin oriented films of silver produced by condensation in a vacuum on cleaved surfaces of single crystals of rocksalt were followed as a function of temperature. Marked agglomeration occurred in films of thicknesses less than 1000A. For thicker films in the 1000–2000A range rectangular holes appeared in the films with their edges aligned with the 110 directions of the rocksalt. The effects are attributed to the readjustment of the oriented surface under the influence of surface tension forces.

THE behavior of thin films of silver condensed in a vacuum on various substrates has been the subject of numerous investigations. A large number of these investigations have been concerned with the measurement of the electrical resistivity and its temperature dependence. The general features of all of the resistance measurements have been discussed by Appleyard.[1] There are two principal results common to all of the resistance measurements. The first is that the resistivity of thin films is greater by factors of 10^3 to 10^7 than that of the bulk metal. The second is the irreversible increase of resistance with increase of temperature. Moreover, for films of a given thickness the resistivity is greater the higher the temperature of the substrate at the time of deposition. These observations led to the view that the film as deposited has a granular structure which is a result of the agglomeration of the metal under the surface tension forces. That such agglomeration occurs has been shown directly by the work of Estermann,[2] and by that of Andrade and Martindale.[3] Recent photographs made with the electron microscope clearly reveal the grained structure of thin metallic films.[4] The energetics and proposed mechanisms of aggregation are given by Appleyard and by Lennard-Jones.[5] It was suggested that aggregation occurs if the attractive forces between the condensed metallic atoms is greater than the forces between the metallic atoms and those of the substrate. The tendency to aggregation is opposed by the periodic field of the surface and aggregation can only occur if the rate of surface diffusion at the given temperature is great enough. It is apparent that an increase of temperature will favor the tendency to aggregation since higher temperature results in more rapid diffusion rates.

In the present work the structural changes of oriented thin films of silver all condensed on rocksalt at 200°C were followed as they were heat-treated subsequently at higher temperature. The films were produced by vacuum distillation of silver on freshly cleaved crystals of rocksalt preheated to about 200 C in accordance with well-known techniques.[6] Wilman and Goche[7] in a detailed study of the structure of such films showed them to possess a twinned structure as grown but heat-treatment at 500 C for a few minutes converted them to relatively perfect single crystals.

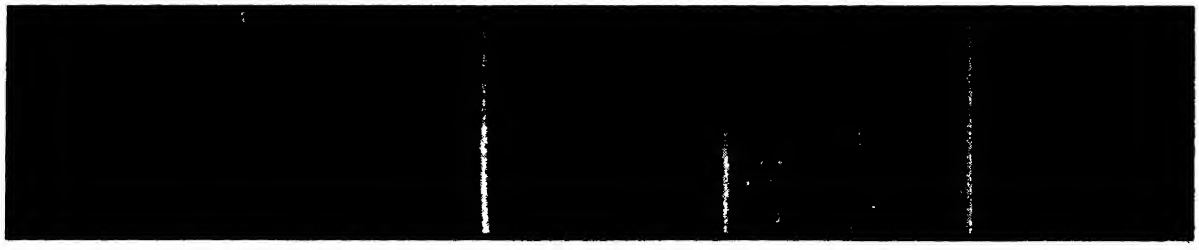
EXPERIMENTAL METHOD

The method of Lassen and Bruck was used to produce thin oriented films of silver. A single crystal of rocksalt was cleaved and was introduced into the tube which was immediately evacuated. At a pressure of less than 0.01 micron and after preheating the rocksalt crystals to about 200 C, spectroscopically pure silver was vaporized from a tungsten filament and condensed on the rocksalt. The vacuum chamber was designed to permit the introduction of five cleaved crystals with dimensions of about 4×4 mm. The crystals were supported on a glass slide movable at will beneath the vaporizer by means of an external magnet. This arrangement made it possible to vary the exposure to the beam of silver. By using different exposure times the film thickness could be varied for the same heating current and for the same geometry.

The films thus produced were examined under vertical illumination with both bright field and polarized light (crossed nicols) using a Bausch and Lomb metallograph. For the films used in this work it was found in general that polarized lighting was more effective in revealing the structure than was bright field illumination. Accordingly most of the photo-micrographs herein reproduced were made under polarized illumination. The electron diffraction attachment of an RCA electron microscope was used to determine the orientation of the films.

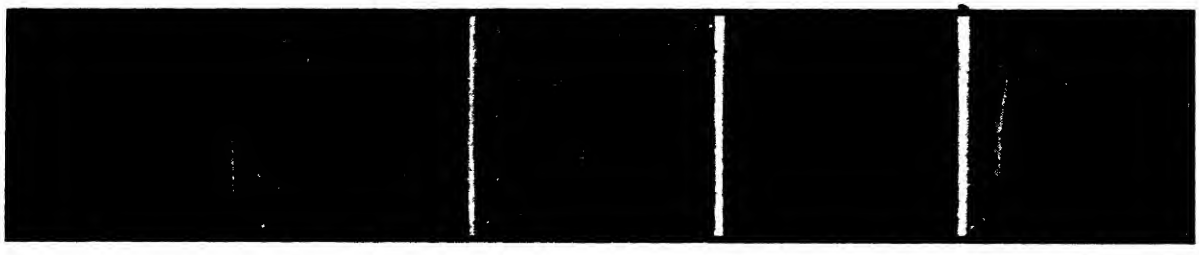
After inspection, the films still supported on the rocksalt substrate were introduced into a vacuum furnace for heat-treatment. The films were heated to various temperatures in the range of 300–500 C for times of from two to twelve hours. After each such

* This work was done under contract with ONR.
† Now at Brookhaven National Laboratory, Upton, New York.
[1] E. T. S. Appleyard, Proc. Phys. Soc. 49, Extra Part, 118 (1937).
[2] I. Estermann, Zeits. f. physik. Chemie 106, 403 (1923).
[3] E. N. da C. Andrade and J. G. Martindale, Phil. Trans. Roy. Soc. 235, 69 (1935).
[4] R. G. Picard and O. S. Duffendack, J. App. Phys. 14, 291 (1943); H. Levinstein, J. App. Phys. 20, 306 (1949).
[5] J. E. Lennard-Jones, Proc. Phys. Soc. 49, Extra Part, 140 (1937).
[6] H. Lassen, Physik. Zeits. 35, 172 (1934); F. Kirchner and H. Lassen, Ann. d. Physik 24, 113 (1935); L. Bruck, Ann. d. Physik 26, 233 (1936); W. Cochrane, Proc. Phys. Soc. 48, 723 (1936); V. O. Rudiger, Ann. d. Physik 30, 505 (1937).
[7] O. Goche and H. Wilman, Proc. Phys. Soc. 51, 625 (1939).



(a) (b) (c) (d) (e)

FIG. 1. Silver film. Thickness~400A. Polarized light crossed nicols 500×. a. As deposited, b. 320°C; 4 hours, c. 400°C; 4 hours, d. 475°C; 4 hours, e. 500°C; 4 hours.



(a) (b) (c) (d) (e)

FIG. 2. Silver film. Thickness~700A. Polarized light crossed nicols 500×. a. As deposited, b. 320°C; 4 hours, c. 400°C; 4 hours, d. 475°C; 4 hours, e. same as (d) but bright field illumination.

treatment the films were examined optically to note any changes in structure. The nominal film thicknesses were calculated from a chemical determination of the silver, using a colorimetric method, the apparent area of the film, and the bulk density. The nominal thickness is the thickness a film would have if it were uniformly distributed in depth over the surface.

RESULTS

Films were prepared in thicknesses extending from 100A to about 4000A. All were heat-treated in a vacuum (less than 0.01 micron) at temperatures between 200 C, the temperature at which they were grown, and 530 C. Above 500 C the rocksalt vaporized profusely with the result that the films were ultimately detached from the substrate and destroyed. It was found that the behavior of the films depended strongly on their thickness as might have been anticipated. For this reason it is convenient to discuss the results in terms of four different ranges of thickness as indicated in the following. All of the films to be described gave diffraction patterns characteristic of single crystals. All photographs reproduced here were made with the cube edges (100 directions) of the rocksalt parallel to the edges of the photographs.

100–500A.—Typical structures of films of this thickness are shown in Fig. 1. These films are noted to have a granular appearance. At higher magnification they are found to possess a network structure which is characteristic of these films and has been observed and reported by many investigators. Recent papers by Levinstein[4] with indium and antimony, and by Picard and Duffendack[4] with aluminum, gold, and copper for unoriented metallic films deposited on collodion, include photo-micrographs of the network structure obtained with an electron microscope. In their work the aggregation of metal into local regions separated by thinner regions or canals is clearly shown. The thinnest films in this thickness range were non-conducting as was indicated when attempts were made to get electron diffraction patterns. Under such conditions it is impossible to get a diffraction pattern because the beam drifts about capriciously as local regions acquire or lose electric charge. The granular structure may be seen on the film as deposited (Fig. 1a). In this case the individual grains are of the order of 1 micron in diameter and on the photo-micrograph (500×) are about 1 mm in diameter. However, the structure becomes accentuated with heat-treatment for a few hours at 320 C (Fig. 1b). There was little observable change in the films on heat-treatment at higher temperatures until a temperature of 500 C was reached, at which point the aggregation of the film became marked (Fig. 1e).

Occasionally the silver as deposited showed a preference for cleavage lines on the crystal. As the temperature was raised the lines faded out. These effects are illustrated in Figs. 1a and 1b which are photographs of the same area before and after heat-treatment.

500–1000A.—The structures observed in this thickness range and their change with temperature can be seen in Fig. 2. The films as deposited resembled those of the thinner range but their changes with temperature are different. On heating to 320 C for four hours the network structure became plainer, with some indication of agglomeration. After heat-treatment at 400 C the

network structure practically disappeared and the silver had collected into relatively large aggregations. At 475 C in some of the films the silver collected along cleavage lines to form ridges and also in small stubby rods extending normal to the surface. The protuberances of silver, whether along cleavage lines or in isolated regions, seemed to grow to about the same height above the crystal surface. The height of growth was about 0.001 mm and was measured by noting the movement of the microscope objective to bring into focus alternately the tops of the growths and the surface of the film.

1000–1500A.—While films of thickness less than 1000A all underwent some adjustment on heating to 320 C, films of greater thickness did not change in appearance until the temperature was increased at least to 400 C. In this thickness range the thinner films on heating to 420 C for four hours were perforated by approximately rectangular windows with the edges aligned with the (110) directions of the rocksalt. The holes were uniformly distributed and were all about the same size within perhaps a factor of two. On increasing the temperature to 450 C the holes increased in size by some 20 percent which was accompanied by the addition of spurs in some cases resulting in changes in shape. In Figs. 3a and 3b are shown some typical holes and also their growth with temperature. Both 3a and 3b were made under bright field illumination and are of the same area of the film. Figure 3c is the same area under polarized light. In this latter photograph the advantage of polarized light for revealing the granular structure is clearly seen, for while the holes are outlined under both bright field and polarized light, the granular structure is only shown under polarized light. For thicker films in this region the rectangular holes appeared also but their density was not as great and the film in general retained its continuity (Figs. 3d and 3e). Further examples of the holes and their growth are shown in Figs. 4a and 4b. Of interest in these photographs is the growth of serrations or steps from a scratch in the film. In Figs. 4c and 4d are shown holes in the neighborhood of a cleaved edge of rocksalt to show the relative orientation of the holes with respect to the (100) directions of the rocksalt. The preference of the straight edges of the windows for the face diagonal directions is clearly illustrated. The density of the windows decreases as a scratch or a free edge is approached. It is of interest also to note that the windows or parts thereof seem to develop from such free edges. (See scratch in Fig. 4b.)

1500–2000A.—Films of this thickness were quite stable and did not show any marked structural changes until heated to 480 C, whereupon the films were found to be criss-crossed by systems of straight lines spaced about 1 micron apart and directed along the (110) directions of the rocksalt. The lines became more pronounced on further heating at 520 C until finally the films disintegrated. Windows sometimes appeared in the form of crosses or rectangular openings with serrated edges. The serrations appeared to have magnitudes corresponding

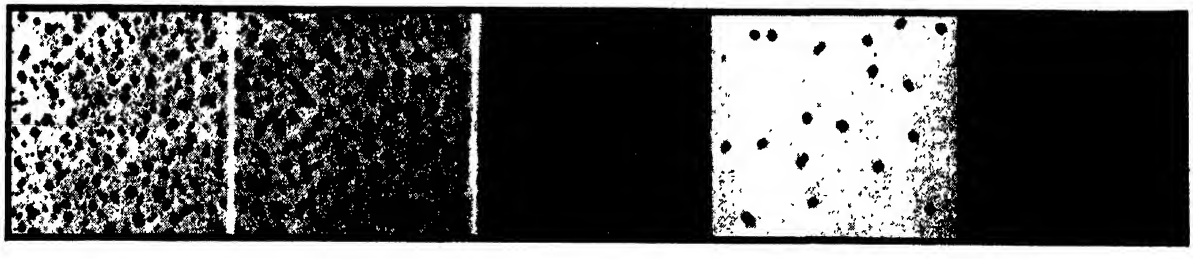


FIG. 3. Silver film. a. Thickness~1100A. 4 hours at 420 C. Bright field 500×, b. same area as (a) after 4 hours at 450 C. Bright field 500×, c. same condition as (b) but polarized light, d. thickness 1400A. 4 hours at 480 C. Bright field 500×, e. same as (d) but polarized light.

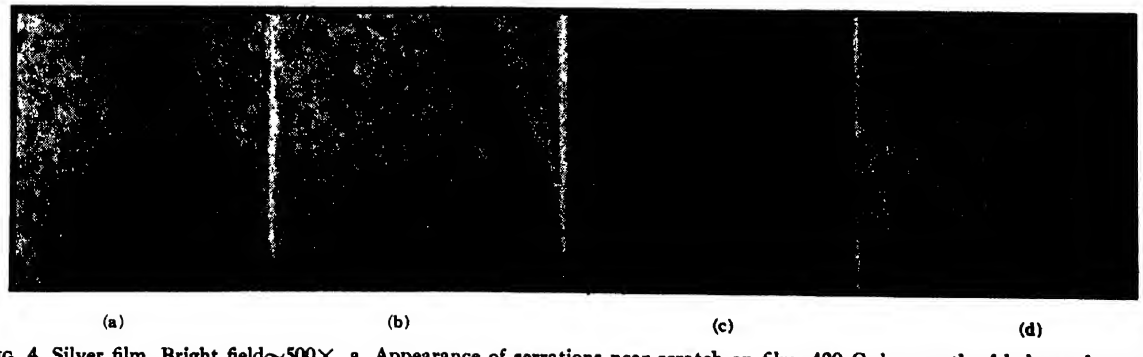


FIG. 4. Silver film. Bright field~500×. a. Appearance of serrations near scratch on film. 420 C, b. growth of holes and serrations at elevated temperature 480 C, c. orientation of holes with respect to cleaved edge of rocksalt substrate 420 C, d. growth of holes at 480 C.

to the observed spacing of the lines covering the entire film. These effects are all shown in Fig. 5. On the original photo-micrographs could be seen systems of faint straight lines running along the face diagonals of the rocksalt.

Above 2000A.—Films thicker than 2000A did not undergo any pronounced structural changes with heat-treatment until they disintegrated after heating to 520 C for eight hours. The heaviest film studied was about 4000A in thickness.

GENERAL OBSERVATIONS

All of the films as deposited gave bright specular reflections. The mirror surfaces were maintained until marked agglomeration occurred or until holes formed in the films at which point the films assumed a matte appearance.

On high rates of deposition the silver would occasionally collect in globules along lines of conchoidal fracture on the cleaved surface of the rocksalt. This effect is illustrated in Fig. 6a.

By exposing the films to a moist atmosphere small sodium chloride crystals could be produced on top of the silver. Such crystals were found to grow over a period of a few months while the films were being stored in a calcium chloride desiccator. They can be readily produced by breathing lightly on a film supported on rocksalt. In Figs. 6b and 6c are shown some of these small crystals. In this case the salt crystals are all aligned with edges parallel to the 110 direction of the substrate and also along the 110 direction of the silver, although often they are rotated 45° with respect to this direction. In the lower right-hand corner of Fig. 6b may be noted one crystallite rotated 45° with respect to the others. These two orientations were the only ones observed on any of the films. Since the silver as deposited was oriented with its structure parallel to that of the rocksalt, then the same relative orientation would be expected if sodium chloride were deposited on silver. The salt on silver should then have the same orientation as the underlying substrate. In the photographs the 100 directions of the rocksalt substrate are parallel to the edges of the photograph. If we assume that the cube edges of the sodium chloride crystallites are the 100 directions then the crystallites are rotated 45° with respect to the expected orientation. This orientation is consistent with that reported by Brück and confirmed by Goche and Wilman for films which were not thoroughly rinsed diffraction points were observed and were attributed to sodium chloride retained on the film with the same orientation as was found here.

To determine whether the epitaxy noted in the foregoing for sodium chloride on an oriented silver film was the result of an influence exerted by the substrate through small pores, a film was floated off the rocksalt in water and picked up on a degreased silver plate.[8] After drying, the film was rinsed in a saturated aqueous solution of sodium chloride and quickly dried. Small crystals of sodium chloride were thus produced on the silver film. Examination under the microscope showed the crystallites to be aligned and with the same orientation as was observed for the films on the rocksalt substrate (Fig. 7). This result is of interest in that it gives a ready means of getting a rough check on the degree of orientation of a

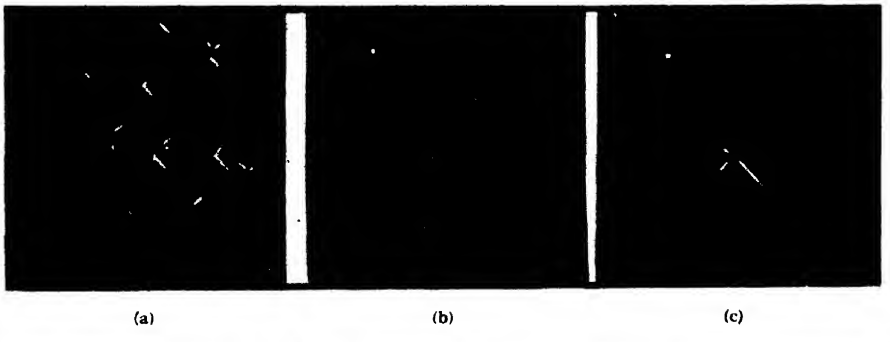


FIG. 5. Silver film. Polarized light (crossed nicols) 500X. a. Thickness 1600A 480 C for 4 hours. b. Thickness 1800A 480 C for 4 hours. c. Same area as (b) after 4 hours at 520 C.

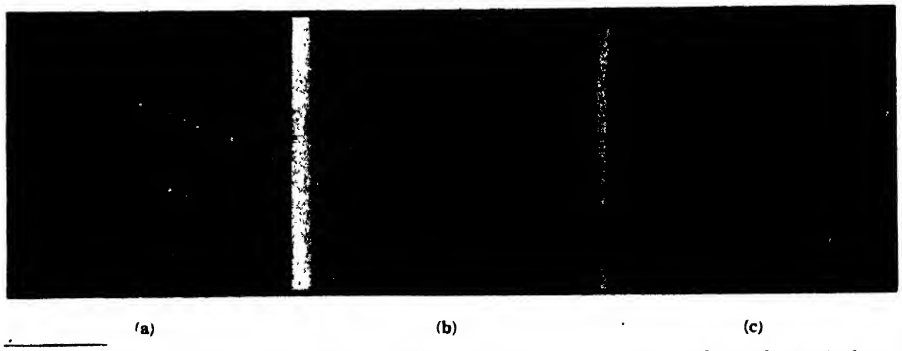


FIG. 6. a. Aggregation of silver on cleavage lines. Polarized light 500X. b. Rocksalt crystals on silver showing epitaxy. Bright field 500X. c. Same as (b). Polarized light 500X.

[8] The author is indebted to Mr. Theodore Watanabe, Washington State College, who worked out the technique of mounting the films.

film without resorting to electron diffraction. Indeed, it may be more sensitive to the orientation in the first layer since the surface forces are probably of short range.

DISCUSSION

The network or granular structure which was observed for the thinner films is characteristic of such films and is well-understood. The metal as deposited agglomerates into small regions under the influence of surface tension forces which for very thin films may be very great. On increasing the temperature further agglomeration may take place. In the present work this effect is noted in the thinnest films where the network structure is accentuated as the temperature is increased and finally the film breaks into small isolated agglomerated regions.

As the film thickness is increased the adjustment of the surface under the effects of the surface tension forces are not as drastic. For films of intermediate thickness this leads to the formation of holes which are evenly distributed and have sides parallel to the 110 direction of the rocksalt. The origin of the holes is suggested in the following. The silver film has the same orientation as the rocksalt and therefore the lines of closest packing of silver atoms are in directions parallel to the face diagonals of the rocksalt. These directions are therefore directions along which the binding between silver atoms is greatest. As the temperature is raised the film is put under increasing tension as a result of the greater thermal expansion of the rocksalt than that of the silver. At some temperature, depending upon the film thickness, the film fails and since it apparently fails along lines of closest packing the process of failure must be one of shear. The surface tension forces then cause the film to contract further. The difference between the coefficients of expansion can only account for about 1 percent total strain. Since the holes are much larger than this, one concludes that the primary cause of growth of the holes lies with the surface tension forces. The holes continue to grow somewhat with increase of temperature. For the thickest films (1500–2000A) the surface tension forces are not sufficiently great to cause large changes in the surface. The film is strained enough, however, to lead to rupture or local yielding sufficient to be visible under the microscope. The lines broaden on further heating indicating an extension of the rupture process but also indicating the films are sufficiently thick to resist any large scale readjustment to the surface tension forces.

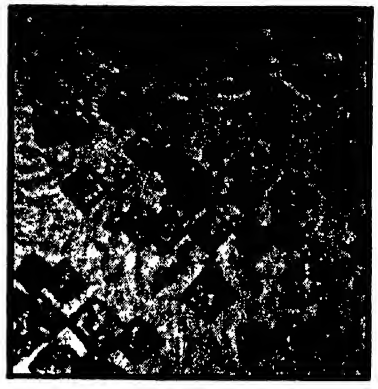


FIG. 7. Sodium Chloride Crystals deposited from an aqueous solution on an oriented silver film supported on a silver plate 400×.

Essig[9] noted in one case the formation of regular holes in a silver film condensed on mica. The windows were bounded by lines intersecting at 60° and 120°. The explanation of the origin of these holes must be similar to that given here for rectangular holes. The silver for the conditions used by Essig must have condensed with the 111 plane parallel to the surface. For this orientation there are three lines of densest packing of atoms in the plane of the surface and they intersect at 120°. Thus when failure occurs if it were by a shearing process it would be expected to take place along lines of closest packing, in which case the windows would have straight sides intersecting at the observed angles. Since the relative orientation of the silver relative to the mica was not determined, it is not known whether the orientations would be consistent with the picture postulated here. However, there appears to be little question about the general description of the result.

ACKNOWLEDGMENTS

The author is indebted to Mr. Mark Adams, Washington State Institute of Technology, for the colorimetric determinations, to Mr. R. F. Smith of Brookhaven National Laboratory, for photo-micrography (Fig. 7), and to Dr. A. W. McReynolds for many helpful criticisms.

[9] S. F. Essig, J. App. Phys. 10, 61 (1939); Phys. Rev. 55, 229 (1939).

Upper and Lower Bounds of Eigenvalues for Composite-Type Regions

E. L. Chu
Microwave Laboratory, Stanford University, California
(Received October 21, 1949)

A survey is made to compare the essential features of different variational methods. Schwinger's method is found to have special merits for boundary value and eigenvalue problems concerning composite-type regions. It can give both an upper bound and a lower bound for the eigenvalue of any mode, one bound being as good as the other. It closely resembles the Courant-Trefftz method and is simpler. In both methods the trial functions are so chosen as to satisfy the differential equation while the boundary or continuity conditions are perturbed. The applicability and power of Schwinger's method, being hitherto demonstrated only for particular examples, are exhibited for a whole class of problems by a general formulation of the method in precise terms. The formulation for the upper-bound case presents no difficulty, and a simple proof is given in this paper. A similar proof for the lower-bound case is possible but has a more restricted field of application. The general formulation places the two different cases on equal footing. A way of avoiding the difficulty of having no exact knowledge of the lower eigenfunctions in higher mode cases is discussed in detail.

I. SURVEY OF METHODS

THE best-known variational method of Rayleigh[1]-Ritz[2] has the great drawback of giving only an upper bound for the true eigenvalue with no information about the accuracy of the bound. One is usually content to indicate the degree of accuracy obtainable by applying the method to problems which can be solved exactly and comparing the approximate value with the true one. This is not quite satisfactory, because the examples chosen may happen to be particularly favorable. Besides, the minimizing process of Ritz is somewhat involved especially for high orders of approximation or for higher modes. There have been recommended, notably by Galerkin,[3] some modifications of the Ritz method, in which the minimizing process is simplified. Yet, the character of overestimation remains the same.

Weinstein[4] has devised an ingenious method which is capable of giving both an upper and a lower bound for the eigenvalue for any one mode. It is rather unfortunate that the expression to be minimized is somewhat complicated, being the difference between a certain integral and the square of another, and the evaluation of one of these integrals is considerably more difficult than that of the Rayleigh-Ritz expression. To facilitate the evaluation of the integrals, only simple trial functions can be used, thus limiting the accuracy of the process. Though interesting and attractive, this method has not been used much in practice.

The suggestion that a lower bound for the eigenvalue can be obtained by calculating the stationary value of the variational expression under loosened boundary conditions or loosened conditions of continuity was first made by Courant.[5] The value calculated must be stationary with respect to small arbitrary variations for otherwise it may not be a lower bound. How this can be done was first shown by Trefftz.[6] Unlike the Rayleigh-Ritz method, the trial functions used by Trefftz satisfy the differential equation but are subject to less restrictive boundary and continuity conditions. The Courant-Trefftz method may be considered as the counterpart of the Rayleigh-Ritz method, because the requirements on the trial functions for these two methods are just opposite to each other. In view of the elegance and the wide applicability of the Courant-Trefftz method, it seems desirable to give here a brief account of its application.

Consider the Sturm-Liouville differential equation for two independent variables x and y:

$$L[u]+\lambda\rho u=(pu_x)_x+(pu_y)_y-qu+\lambda\rho u=0, \quad (p>0,\ \rho>0), \quad (1)$$

for a region R with boundary Γ.[7] The boundary condition is given as $u=f(s)$. According to Trefftz, the problem is to be modified by replacing the given boundary condition $u=f(s)$ with a set of weaker ones:

$$\int_\Gamma \{w-f(s)\}g_i(s)ds=0, \quad i=1, 2, \cdots, m,$$

where $g_i(s)$ are arbitrary functions chosen in such a way that the differential equation $L[w]+\mu\rho w=0$ can be solved exactly. As in the Rayleigh-Ritz method, the function w that corresponds to the lowest eigenvalue μ will minimize the expression

$$D[w]=\int_R\int p(w_x^2+w_y^2)dxdy+\int_R\int qw^2dxdy, \quad (2)$$

but under a set of modified boundary conditions. The

[1] Lord Rayleigh, Phil. Trans. **161**, 77 (1870); *Scientific Papers*, (Cambridge University Press, London, 1899), Vol. 1, p. 57.
[2] W. Ritz, *Oeuvres* (Gauthier-Villars, Paris, 1911), pp. 192–316.
[3] See F. Pfeiffer, Handbuch der Physik **6**, 345 (1928); also Biezeno-Grammel, Tech. Dynamik (Verlag. Julius Springer, Berlin, 1939), 167–169 (1939).
[4] D. H. Weinstein, Proc. Nat. Acad. Sci. **20**, 529 (1934); see also J. K. L. MacDonald, Phys. Rev. **46**, 828 (1934).
[5] R. Courant, Math. Ann. **97**, 711 (1927).
[6] E. Trefftz, Math. Ann. **100**, 503–521 (1928).
[7] In Eq. (1), p, q, and ρ are functions of x and y while λ is a constant. $p>0$ and $\rho>0$ in R. $L[u]=(pu_x)_x+(pu_y)_y-qu$ is simply a definition.

normalization relation

$$\int_R\int \rho w^2 dxdy = 1$$

is understood.

Since the modified conditions are weaker than the given condition, that is, the set of compatible functions for w contains the set of functions for u, so the minimum value of $D[w]$ can never be greater than that of $D[u]$. Thus $\mu \leqq \lambda$, μ, and λ are the minimum values of $D[w]$ and $D[u]$, respectively.

The modified problem is the same as that of finding a function w for which the following functional $I[w]$ is stationary:

$$I[w] = D[w] - \mu\int_R\int \rho w^2 dxdy - 2\sum_1^m c_i \int_\Gamma \{w - f(s)\} g_i(s) ds, \quad (3)$$

where c_i's are Lagrangian multipliers. This can be shown by putting the first variation of $I(w)$ equal to zero, to obtain

$$\int_R\int \{L[w] + \mu\rho w\}\delta w dxdy - \int_\Gamma \left\{p\frac{\partial w}{\partial n} - \sum_1^m c_i g_i(s)\right\}\delta w ds = 0.$$

The volume integral vanishes if w satisfies the equation $L[w] + \mu\rho w = 0$. The surface integral will also vanish for any arbitrary δw if

$$p\frac{\partial w}{\partial n} = \sum_1^m c_i g_i(s)$$

on Γ. These requirements on w can be fulfilled if w is taken as

$$w = \sum_1^m c_i w_i(x, y)$$

and $g_i(s)$ as

$$g_i(s) = p(\partial w_i(x, y)/\partial n),$$

where the w_i's are any m particular solutions of the differential equation $L[w] + \mu\rho w = 0$.

The $m+1$ constants, μ and the c_i's, are to be determined from $m+1$ simultaneous equations, m of which are the above boundary conditions while the other one is the normalizing relation. As shown above, μ is a lower bound for λ.

The lower bound may be improved by imposing more stringent boundary conditions. In fact, as m approaches infinity, w may be made to approach the true eigenfunction. Thus μ will approach λ. In some cases it is desirable not to relax the boundary conditions but the conditions of continuity. Such cases have also been considered by Trefftz.

In passing, we may remark that the Trefftz method can also be used for calculating upper bounds. What we have to do is simply to make the boundary conditions more stringent than given in the original problem. However, there would be no advantage to do so even where the trial functions can be easily chosen, because the algebraic process of the Trefftz method is still more involved than that of Ritz.

One method other than Ritz's for calculating upper bounds needs special mentioning. This was described by Grammel[8] and may be considered as the counterpart of Ritz's method or rather that of Galerkin's method, from the viewpoint that this method is more intimately connected with the integral equation rather than with the differential equation.

To be specific, if the differential and integral equations for the same problem are given as $L[u] + \lambda\rho u = 0$ and

$$u(x, y) = \lambda\int_R\int G(x, y; x', y'; 0) \times u(x', y')\rho(x', y')dx'dy' \quad (4)$$

respectively, where $G(x, y; x', y'; 0)$ is the appropriate Green's function of the equation $L[u] = 0$, then the Grammel method deals with the equation

$$\int_R\int \left[w(x, y) - \lambda\int_R\int G(x, y; x', y'; 0)w(x', y') \times \rho(x', y')dx'dy'\right]\rho(x, y)\delta w dxdy = 0 \quad (5)$$

while Galerkin's method deals with

$$\int_R\int \{L[w] + \lambda\rho w\}\delta w dxdy = 0. \quad (6)$$

In both methods we take

$$w(x, y) = \sum_1^m c_i\phi_i(x, y) \quad (7)$$

as the trial function. The functions ϕ_i are any suitable coordinate functions satisfying the appropriate boundary conditions of the problem. We restrict ourselves only to such variations as

$$\delta w = \sum_1^m \phi_i(x, y)\delta c_i.$$

The variational problem is then reduced to the solution

[8] R. Grammel, Proc. 5th Intern. Congress for App. Mech. 691–693 (1938); also Biezeno-Grammel, reference 3.

of m simultaneous equations. They are

$$\sum_1^m c_i \int_R\int\Big[\phi_i(x,y)-\lambda\int_R\int G(x,y;x',y';0)\times\phi_i(x',y')\rho(x',y')dx'dy'\Big]\times\phi_k(x,y)\rho(x,y)dxdy=0,\quad k=1,2,\cdots,m \quad (8)$$

in Grammel's method and

$$\sum_1^m c_i\int\int\{L[\phi_i]+\lambda\rho\phi_i\}\phi_k dxdy=0,\quad k=1,2,\cdots,m \quad (9)$$

in Galerkin's.

The former set of equations appear to be more formidable but actually they are more tractable and easier for calculation than the latter set. This is due to the fact that the choice of ϕ_i's in Grammel's method is always easier than in Galerkin's, because, there, no differentiations on ϕ_i's are involved. If the same expression, i.e., the same linear combination of the same ϕ_i's is used as the trial function for both cases, Grammel has shown that the upper bound calculated by his method is always lower than that calculated by Galerkin's or Ritz's method, the latter two giving exactly the same value.

In some cases we do not even need to know the explicit expression of the Green's function for carrying out the Grammel process. It is usually possible to transform the double integral into a single one

$$\int_R\int\phi_k(x,y)\rho(x,y)dxdy\times\int_R\int G(x,y;x',y';0)\phi_i(x',y')\rho(x',y')dx'dy'=\int_R\int\frac{M(\phi_i,\rho)\cdot M(\phi_k,\rho)}{\alpha(x,y)}dxdy,$$

where M and α can in some cases be easily recognized as known physical quantities.[9] Thus, the set of simultaneous equations may be reduced to

$$\sum_1^m c_i\Big[\int_R\int\phi_i\phi_k\rho dxdy-\lambda\int_R\int\frac{M(\phi_i,\rho)\cdot M(\phi_k,\rho)}{\alpha}dxdy\Big]=0,\quad k=1,2,\cdots,m.$$

[9] For example, M may represent the bending moment in dynamics, then α will be the modulus of rigidity. See R. Grammel and Biezeno-Grammel, references 3 and 8.

In all the methods discussed so far, we need to set up a set of simultaneous equations and to solve them. The process is usually lengthy and difficult. Besides, the accuracy for higher modes is much worse than for the fundamental, because the values for different modes all come out of the same set of equations. No use has been made of the fact that a given eigenfunction is orthogonal to all the lower ones.

If we know approximately the first $n-1$ eigenfunctions, the knowledge can be put into immediate use for calculating the lower bound for the next higher eigenvalue by the help of the so-called maximum-minimum-property of the eigenvalues.[10] We first consider the problem with the original boundary condition $u=f(s)$ but the loosened orthogonality relations

$$\int_R\int w\cdot v_i\rho dxdy=0,\quad i=1,2,\cdots,n-1,$$

where w is the trial function and v_i's are the first $n-1$ approximate eigenfunctions. The would-be eigenvalue for this perturbed problem will be less than the true value on account of its maximum-minimum-property. Then if we also loosen the boundary condition in a similar way as in Trefftz's method, the calculated value will be still lower than the would-be eigenvalue. The more we know about the v_i's, the better will be the calculated lower bound. However, the process is not quite so simple as briefly described above and cannot be applied for calculating upper bounds.

From the above discussion, it is to be expected that some modification of the Courant-Trefftz method formulated on the basis of the integral equation will have the merit of being both simple and flexible. In fact, such has already been furnished by Schwinger.[11] His method for calculating both upper and lower bounds proves to be very successful. It embodies Courant's idea of perturbing boundary conditions and looks like Grammel's method in using an integral equation.

Schwinger has only worked on boundary value problems, while his method can be used for the calculation of eigenvalues as well. We have done some calculations by his method on the resonant frequency of a disk-loaded cavity and have described in a previous paper[12] our results and how we proceeded with our calculations. We have explained how this method can give an upper bound in one type of calculation and a lower bound in another, one being almost as good as the other.

Despite the stringent condition that the trial functions are to satisfy the differential equation, the method can be applied to quite a large number of important physical problems. It is particularly useful for those

[10] For a proof of the property, see Courant u. Hilbert, *Methoden der Mathematischen Physik* (Verlag. Julius Springer, Berlin, 1931), Vol. I, p. 351–353.

[11] J. Schwinger, "Discontinuities in wave guides" (lecture notes prepared by D. S. Saxon).

[12] E. L. Chu and W. W. Hansen, J. App. Phys. 20, 280–285 (1949).

problems concerning composite-type regions. A region is called composite, if it can be divided by simple surfaces into two or more regions, in each of which the problem can be exactly solved once we impose simple enough conditions on those dividing surfaces. The disk-loaded cavity, which we have considered, is one example.

To facilitate further discussions, it is desirable to give a brief summary of the method in simple concise terms. We take $G(x, y; x', y'; \lambda)$ as the Green's function of the differential equation $L[u]+\lambda\rho u=0$, such that either

$$G=0 \text{ on } \Gamma \tag{10a}$$

or

$$\partial G/\partial n=0 \text{ on } \Gamma. \tag{10b}$$

The integral equation to be solved will be either

$$\frac{\partial u}{\partial n}(s)=-\sigma u(s)=-\int_\Gamma pu\frac{\partial}{\partial n}\frac{\partial G(\lambda)}{\partial n'}ds' \tag{11a}$$

or

$$u(s)=-\frac{1}{\sigma}\frac{\partial u}{\partial n}(s)=\int_\Gamma pG(\lambda)\frac{\partial u}{\partial n'}ds' \tag{11b}$$

as the case may be. $\sigma=-(1/u)(\partial u/\partial n)$ is known as given boundary condition and is usually a piecewise continuous function of s. These equations are not exactly soluble, so we perturb the problem and solve $L[w]+\mu\rho w=0$ with boundary values related by

$$\frac{\partial w}{\partial n}(s)=-\xi w(s)=-\int_\Gamma pw\frac{\partial}{\partial n}\frac{\partial G(\mu)}{\partial n'}ds' \tag{12a}$$

or

$$w(s)=-\frac{1}{\xi}\frac{\partial w}{\partial n}(s)=\int_\Gamma pG(\mu)\frac{\partial w}{\partial n'}ds', \tag{12b}$$

where ξ may be left undefined or chosen properly. If we demand

$$\int pw\frac{\partial w}{\partial n}ds=\int_\Gamma p\frac{\partial u}{\partial n}\cdot\frac{w^2}{u}ds$$

or

$$\int pw\frac{\partial w}{\partial n}ds=\int_\Gamma p\frac{u}{\partial u/\partial n}\cdot\left(\frac{\partial w}{\partial n}\right)^2ds,$$

we obtain

$$\int_\Gamma p\frac{\partial u}{\partial n}\frac{w^2}{u}ds=-\int_\Gamma pwds\int_\Gamma pw\frac{\partial}{\partial n}\frac{\partial G(\mu)}{\partial n'}ds' \tag{13a}$$

or

$$\int_\Gamma p\frac{u}{\partial u/\partial n}\left(\frac{\partial w}{\partial n}\right)^2ds=\int_\Gamma p\frac{\partial w}{\partial n}ds\int pG(\mu)\frac{\partial w}{\partial n'}ds'. \tag{13b}$$

By assuming certain functions for $w(s)$ or $(\partial w/\partial n)(s)$, we can calculate μ from (13a) or (13b) respectively. In both types of calculation we may approach the limiting case $w=u$ and $\mu=\lambda$ as far as may be desired.

The value calculated is only an approximation to the true eigenvalue. The accuracy depends on how we perturb the boundary condition, that is, how we choose w. If we have a good approximate knowledge about u to start with, we may be able to get a very accurate value by a single trial without dealing with the simultaneous equations or solving a determinantal one.

Whether the calculated value is an upper or a lower bound depends on whether we have strengthened or weakened the supplementary conditions of the problem. Except for very simple cases, the distinction between a stringent and a weak condition is not quite obvious. Even where we are sure we have weakened the supplementary conditions, the existence of a lower bound depends on a further condition that the calculated value should be stationary with respect to any small variation in all the admissible trial functions. In the simple process described above we have not gone through such variational considerations.

Schwinger has discussed such questions on boundary value problems. He has shown that his variational expressions possess the property of being definite. Yet, further justifications are necessary in order to formulate some general rules for the calculation of eigenvalues. It is the purpose of this paper to concern ourselves with such justifications.

II. THE FUNDAMENTAL MODE

Let λ_1 and u_1 be the lowest eigenvalue and the corresponding eigenfunction for the Sturm-Liouville differential Eq. (1). u_1 will have no zeros in the region R. It is to be found by minimizing the expression

$$J[u]=D[u]+\int_\Gamma p\sigma u^2ds, \tag{14}$$

where $D[u]$ is as given in Eq. (2), under the boundary condition that $u=0$ at least on that part of Γ where u_1 is zero, and also under the normalizing condition

$$\int_R\int\rho u^2dxdy=1.$$

u should be continuous and have piecewise continuous first derivatives in R, but is otherwise arbitrary. σ is any given piecewise continuous function of coordinates on Γ, such that

$$(\partial u_1/\partial n)+\sigma u_1=0 \text{ on } \Gamma. \tag{15}$$

The expression to be minimized may be written as

$$J[u]=D[u]-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds. \tag{16}$$

This is sensible because $u=0$ wherever $u_1=0$ on Γ. The expression (16) may be transformed by Green's theorem, to obtain

$$D[u]-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds$$
$$=D[u]-D\left[u_1,\frac{u^2}{u_1}\right]-\int_R\int\frac{u^2}{u_1}L[u_1]dxdy,$$

where

$$D[u, v]=\int_R\int p(u_xv_x+u_yv_y)dxdy+\int_R\int quvdxdy.$$

Thus

$$D[u]-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds=\int_R\int \frac{p}{u_1^2}(u_1\nabla u-u\nabla u_1)^2dxdy - \int_R\int \frac{u^2}{u_1}L[u_1]dxdy.$$

Since

$$L[u_1]+\lambda_1\rho u_1=0,$$

we get

$$D[u]-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds = \int_R\int \frac{p}{u_1^2}(u_1\nabla u-u\nabla u_1)^2dxdy + \lambda_1\int_R\int \rho u^2dxdy. \quad (17)$$

In carrying out the above transformation, it is tacitly assumed that u has first derivatives everywhere in R. If not, we may consider R as being composed of several regions in each of which u is everywhere differentiable. Equation (17) brings out the variational principle explicitly in view. It states that λ_1 is the minimum value of $J[u]$ for any admissible u that is properly normalized, because

$$\int_R\int \frac{p}{u_1^2}(u_1\nabla u-u\nabla u_1)^2dxdy\geqq 0.$$

Now we restrict ourselves to use only those trial functions which satisfy the differential equation $L[u]+\lambda\rho u=0$. From (17) and

$$D[u]=-\int_R\int uL[u]dxdy+\int_\Gamma pu\frac{\partial u}{\partial n}ds = \lambda\int_R\int \rho u^2dxdy+\int_\Gamma pu\frac{\partial u}{\partial n}ds,$$

we obtain

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds = \int_R\int \frac{p}{u_1^2}(u_1\nabla u-u\nabla u_1)^2dxdy - (\lambda-\lambda_1)\int_R\int \rho u^2dxdy. \quad (18)$$

If we demand

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds-\int_\Gamma p\frac{u^2}{u_1}\frac{\partial u_1}{\partial n}ds\leqq 0,$$

then

$$\lambda\geqq\lambda_1.$$

Thus we formulate a simple rule for calculating an upper bound for λ_1 as follows:

Choose a function $u(s)$ such that $u(s)=0$ on Γ wherever $\sigma=\infty$. Derive $\partial u/\partial n$ from $u(s)$ such that $u(x, y)$ satisfies $L[u]+\lambda\rho u=0$. Solve the equation

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds=-\int_\Gamma p\sigma u^2ds \quad (19)$$

for λ. The value obtained is always an upper bound.

The method of solution is as follows:

Since

$$u(x, y)=-\int_\Gamma pu\frac{\partial G}{\partial n'}ds',$$

the derived expression for $\partial u/\partial n$ is

$$\frac{\partial u}{\partial n}=-\int_\Gamma pu\frac{\partial}{\partial n}\frac{\partial G}{\partial n'}ds',$$

where $G(x, y; x', y'; \lambda)$ is the Green's function of the equation $L[u]+\lambda\rho u=0$, such that $G=0$ on Γ.[13] Thus

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds=-\int_\Gamma puds\int_\Gamma pu\frac{\partial}{\partial n}\frac{\partial G}{\partial n'}ds' = \int_\Gamma\int_\Gamma p(s)u(s)K(s, s'; \lambda)p(s')u(s')dsds', \quad (20)$$

with

$$K(s, s'; \lambda)=-\left[\frac{\partial}{\partial n}\frac{\partial}{\partial n'}G(x, y; x', y'; \lambda)\right]_{\substack{x, y \text{ on } \Gamma \\ x', y' \text{ on } \Gamma.}}$$

and (19) becomes

$$\int_\Gamma\int_\Gamma p(s)u(s)K(s, s'; \lambda)p(s')u(s')dsds'=-\int_\Gamma p\sigma u^2ds.$$

In some cases it is easier to find $K(s, s'; \lambda)$ than it is to find G and then differentiate it. From this equation λ can be determined when $u(s)$ is chosen. λ depends, of course, on $u(s)$ but the above rule implies that this equation can never be satisfied with $\lambda<\lambda_1$ for any admissible $u(s)$.

The function $u(s)$ should be a good approximation to

[13] It is expedient to have $G=0$ only on that part of Γ, say Γ', where the boundary condition is perturbed and $\partial G/\partial n+\sigma G=0$ on the rest of Γ. Then G is appropriate on $\Gamma-\Gamma'$ and $u(x, y)=-\int_{\Gamma'} pu(\partial G/\partial n')ds'$.

$u_1(s)$ and should be so chosen that the integration of $\int_\Gamma pu(\partial u/\partial n)ds$ will be easy. With a good choice of $u(s)$, which is usually practicable, the derived function $u(x, y)$ will be a very good approximation to $u_1(x, y)$, so the calculated value of λ can be very accurate. The Green's function G or the kernel K is usually expressed in the form of an infinite series; we need to sum it before we can solve the Eq. (19). The algebraic process of summing a series, though sometimes difficult, is much simpler than the solution of a high order determinant.

To explain how to obtain a lower bound for the eigenvalue, we interchange u with u_1 and λ with λ_1 in Eq. (18), and thus obtain an equation which is the mate of (18):

$$\int_\Gamma pu_1\frac{\partial u_1}{\partial n}ds-\int_\Gamma p\frac{u_1^2}{u}\frac{\partial u}{\partial n}ds$$

$$=\int_R\int\frac{p}{u^2}(u\nabla u_1-u_1\nabla u)^2dxdy$$

$$-(\lambda_1-\lambda)\int\int\rho u_1^2dxdy. \quad (21)$$

Here we must have the condition that $u_1=0$ on Γ wherever $u=0$ on Γ'. In other words, u can be zero on Γ only where u_1 is known to be zero. Furthermore, u can have no zeros inside R. The latter restriction implies that u must be the solution for the fundamental mode of the modified problem.

If we demand

$$\int_\Gamma pu_1\frac{\partial u_1}{\partial n}ds-\int_\Gamma p\frac{u_1^2}{u}\frac{\partial u}{\partial n}ds\leqq 0, \quad (22)$$

we should have $\lambda\leqq\lambda_1$. But this is not as practicable as the upper-bound case, because u_1 and $u_1(\partial u_1/\partial n)$ are not supposed to be known simultaneously on Γ except where $u_1=0$. The boundary condition is given by $(1/u_1)(\partial u_1/\partial n)$, not by both u_1 and $\partial u_1/\partial n$. Thus we have to demand $(1/u)(\partial u/\partial n)\geqq(1/u_1)(\partial u_1/\partial n)$ everywhere on that part of Γ where $u_1\neq 0$ in order that the condition (22) will be certainly satisfied. If we are going to perturb the boundary condition slightly, the modified problem may not be exactly soluble; if we perturb more, the calculated value may not be accurate enough.

One important case, however, can be dealt with immediately. It occurs most often in practical problems especially those concerning mechanical and electromagnetic oscillations. In this case, the boundary condition is given as $u_1(\partial u_1/\partial n)=0$ everywhere on Γ; so if we choose u such that $\partial u/\partial n=0$ wherever $u_1\neq 0$ (i.e., $\partial u_1/\partial n=0$) on Γ, we will have $\lambda\leqq\lambda_1$.

In general, it is possible to formulate a rule similar to that for the upper-bound case. The rule is as follows:

Choose a function $(\partial u/\partial n)(s)$ *such that* $\partial u/\partial n=0$ *on* Γ *wherever* $\sigma=0$. *Derive* $u(s)$ *from* $\partial u/\partial n$ *such that* $u(x, y)$ *satisfies* $L[u]+\lambda\rho u=0$. *Solve the equation*

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds=-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial u}{\partial n}\right)^2ds \quad (23)$$

for λ. *The value obtained is always a lower bound.*

It must be understood here that $\partial u/\partial n$ should be reasonably close to $\partial u_1/\partial n$, otherwise we might get a solution for a higher mode.

The solution is obtained in a similar way as in the upper-bound case.

Here $u(x, y)=\int_\Gamma pG(\partial u/\partial n')ds'$, in which $G(x, y; x', y'; \lambda)$ is the Green's function of $L[u]+\lambda\rho u=0$ such that $(\partial G/\partial n)=0$ on Γ.[14] Thus

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds=\int\int_\Gamma p(s)\frac{\partial u}{\partial n}G(s, s'; \lambda)p(s')\frac{\partial u}{\partial n'}dsds'. \quad (24)$$

The above rule implies that the equation

$$\int_\Gamma\int_\Gamma p(s)\frac{\partial u}{\partial n}G(s, s'; \lambda)p(s')\frac{\partial u}{\partial n'}dsds'=-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial u}{\partial n}\right)^2ds$$

can never be satisfied with $\lambda>\lambda_1$ for any admissible trial function.

The proof of the rule, however, is not so easy as for the upper-bound case. We are not going to be concerned with it here, since it is but a special case of that for the higher modes, for which a general proof is given below.

So far, we have only considered the perturbation of boundary conditions. Similar results will be obtained by perturbation of the conditions of continuity. Discussion of the latter case will also be reserved for the higher modes.

III. HIGHER MODES

1. Upper Bound

We consider the lth mode. Let λ_l and u_l be the lth eigenvalue and the corresponding eigenfunction. u_l has zeros (nodes or nodal lines or nodal surfaces) inside R; they divide R into no more than l sub-regions.[15] It is to be found by minimizing the same expression (14) as for the fundamental mode. The supplementary conditions as to boundary values, the normalizing condition and the continuity requirements are also the same, but there are $l-1$ additional orthogonality conditions. They are:

$$\int\int\rho uu_i dxdy=0, \quad i=1, 2, \cdots, l-1. \quad (25)$$

For the case of a higher mode, the variational expression cannot be transformed into a form as represented by Eq. (17), except when u and u_l have exactly the same zeros inside R. In general, such transformation

[14] See reference 13. If $(\partial G/\partial n)=0$ on Γ' and G is appropriate on $\Gamma-\Gamma'$, then $u(x, y)=\int_{\Gamma'}pG(\partial u/\partial n')ds'$.

[15] Courant u. Hilbert, reference 10, S. 393.

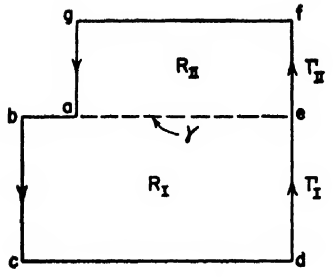


FIG. 1. A composite region R is shown divided into two simple regions R_I and R_{II} by the surface γ. The open boundary *abcde* is denoted by Γ_I and *efga* by Γ_{II}.

cannot be justified. We may, of course, cut out those singularities and then transform the expression, but the result so obtained does not appear to have an explicit extremal property.

It is easier to proceed directly from the variational principle. With u satisfying all the appropriate supplementary conditions mentioned above, the principle assures us that the expression has a minimum value equal to λ_1. The minimum is reached when and only when $u=u_1$. Thus we have, for any admissible u,

$$D[u]+\int_\Gamma p\sigma u^2 ds \geqq \lambda_1 \int_R\int \rho u^2 dxdy. \qquad (26)$$

If u satisfies the differential equation $L[u]+\lambda\rho u=0$, the inequality (26) may be transformed to

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds+\int_\Gamma p\sigma u^2 ds \geqq (\lambda_1-\lambda)\int_R\int \rho u^2 dxdy. \qquad (27)$$

If we demand, as before,

$$\int_\Gamma pu\frac{\partial u}{\partial n}ds+\int_\Gamma p\sigma u^2 ds \geqq 0,$$

then

$$\lambda \geqq \lambda_1.$$

The rule for calculating an upper bound for a higher eigenvalue is just the same as for the lowest except that the trial functions must be orthogonal to all the lower eigenfunctions. If we really need to know the exact form of these latter functions, the method would not be of any practical use. For problems which cannot be solved exactly, we have, at most, only an approximate knowledge about these lower eigenfunctions. If we use approximate functions instead of true ones, the calculated value will be somewhat lower than it would be according to the maximum-minimum-property of eigenvalues. Thus we are not sure whether it is really an upper bound. With the property that u satisfies the differential equation, as it must in this method, this kind of difficulty can easily be removed. We will show how this can be done in a later section.

Now we come to the question of the perturbation of continuity conditions. Let us divide the region R into R_I and R_{II} by γ, as shown in Fig. 1. Let Γ_I and Γ_{II} denote those parts of Γ which belong to R_I and R_{II}, respectively. We choose a function $u_I(s)$ on $\Gamma_I+\gamma$ and $u_{II}(s)$ on $\Gamma_{II}+\gamma$ with $u_I=u_{II}$ on γ, and then derive the trial functions $u_I(x, y)$ and $u_{II}(x, y)$ for the two regions such that they both satisfy the same differential equation.

Thus, u is continuous everywhere in R but the normal component of ∇u is in general discontinuous on γ. u is to satisfy all those restrictions imposed on the trial function for a higher mode for the case of perturbation of boundary conditions, but has no restriction on γ except that it should be continuous there. From the inequality (26), we have instead of (27)

$$\int_{\Gamma_I+\gamma} pu_I\frac{\partial u_I}{\partial n}ds+\int_{\Gamma_{II}+\gamma} pu_{II}\frac{\partial u_{II}}{\partial n}ds + \int_{\Gamma_I+\Gamma_{II}} p\sigma u^2 ds \geqq (\lambda_1-\lambda)\int_R\int \rho u^2 dxdy. \qquad (28)$$

The integrals

$$\int_{\Gamma_I+\gamma} pu_I\frac{\partial u_I}{\partial n}ds \quad \text{and} \quad \int_{\Gamma_{II}+\gamma} pu_{II}(\partial u_{II}/\partial n)ds$$

can be expressed as similar functions of $u_I(s)$ and $u_{II}(s)$ as shown by Eq. (20). We obtain an upper bound for λ_1 by solving the following equation

$$\int_{\Gamma_I+\gamma} pu\frac{\partial u}{\partial n}ds+\int_{\Gamma_{II}+\gamma} pu\frac{\partial u}{\partial n}ds=-\int_\Gamma p\sigma u^2 ds.$$

2. Lower Bound

We perturb the problem by replacing the boundary condition $\partial u_1/\partial n+\sigma u_1=0$ with $\partial w/\partial n+\xi w=0$ such that w does not have to be zero on that part of Γ where $u_1=0$ and $\xi(s)$ is finite wherever $\sigma(s)$ is finite. $\xi(s)$ is a piecewise continuous function and when chosen is to be considered as fixed. The modified problem is to minimize the expression

$$J_1[v]=D[v]+\int_\Gamma p\xi v^2 ds \qquad (29)$$

under the same orthogonality, normalizing and continuity conditions as those imposed on u for minimizing the expression

$$J[u]=D[u]+\int_\Gamma p\sigma u^2 ds$$

of the original problem. Only the boundary conditions are different. Here $v(s)=0$ wherever $\xi(s)$ is infinite, i.e., $w(s)=0$ and $v(s)$ may not be zero where $u_1(s)=0$.

Let us denote the set of trial functions for the original problem by u and that for the modified problem by v. It is obvious that every function of u is contained in v, so u_1 is admissible as a trial function for the expression (29). Let w be that function of the set v which minimizes $J_1[v]$ and μ its minimum value, i.e., $J_1[v]_{\min}=J_1[w]$

$=\mu$. According to the usual variational principle w must satisfy the differential equation $L[w]+\mu\rho w=0$ besides the boundary condition $\partial w/\partial n+\xi w=0$.

Since $J_1[u_l]\geqq J_1[w]$, we obtain easily

$$\int pu_l\frac{\partial u_l}{\partial n}ds+\int_\Gamma p\xi u_l^2ds\geqq\mu\int_R\rho w^2dxdy-\lambda_l\int\rho u_l^2dxdy,$$

i.e.,

$$\int_\Gamma pu_l\frac{\partial u_l}{\partial n}ds-\int_\Gamma p\frac{u_l^2}{w}\frac{\partial w}{\partial n}ds\geqq\mu-\lambda_l. \tag{30}$$

Thus, if

$$\int_\Gamma pu_l\frac{\partial u_l}{\partial n}ds-\int_\Gamma p\frac{u_l^2}{w}\frac{\partial w}{\partial n}ds\leqq 0,\quad \mu\leqq\lambda_l \tag{31a}$$

and conversely if

$$\mu\geqq\lambda_l,\quad \int_\Gamma pu_l\frac{\partial u_l}{\partial n}ds-\int_\Gamma p\frac{u_l^2}{w}\frac{\partial w}{\partial n}ds\geqq 0. \tag{31b}$$

The inequality (30) is the generalization of (21) for higher modes.

As mentioned before in connection with the fundamental mode, the condition (31a) is not quite practical because u_l and $u_l(\partial u_l/\partial n)$ are not known simultaneously on Γ except for certain special cases. In fact, (31a) is not a necessary condition for $\mu\leqq\lambda_l$ and there may exist other conditions which can better serve the purpose. To find one of more general usefulness, it is natural to think of replacing u_l in (31a) by w. One way of doing this is to demand instead of (31a)

$$\int_\Gamma p\xi w^2ds-\int_\Gamma p\sigma w^2ds=0,$$

but this necessitates that $w(s)=0$ wherever σ is infinite and is only soluble for $\mu\geqq\lambda_l$.

Alternatively we may demand the following condition

$$\int_\Gamma p\xi w^2ds-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2ds=0,$$

i.e.,

$$\int\frac{p}{\xi}\left(\frac{\partial w}{\partial n}\right)^2ds-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2ds=0 \tag{32a}$$

with $\partial w/\partial n=0$ on Γ wherever $\sigma=0$. Since w satisfies the differential equation $L[w]+\mu\rho w=0$, the above condition may be written as

$$-\int_\Gamma p\frac{\partial w}{\partial n}ds\int p\frac{\partial w}{\partial n'}G(s,s';\mu)ds'-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2ds=0. \tag{32b}$$

Here $G(x,y;x',y';\mu)$ is the Green's function of $L[w]+\mu\rho w=0$ such that $\partial G/\partial n=0$ on Γ. It will be shown immediately that the condition (32a), i.e., (32b) is indeed sufficient to make $\mu\leqq\lambda_l$ for any admissible trial function used for $\partial w/\partial n$.

To prove this, we first note that if $\xi=\sigma$ except where $\xi=0$ but $\sigma>0$, (31a) is certainly satisfied and (32a) becomes an identity. For any such ξ we have a unique solution with $\mu\leqq\lambda_l$. In other words, with such values of μ we can always find a function $\partial w/\partial n$ which satisfies (32b). However, we need to prove the converse statement, i.e., if $\mu>\lambda_l$ (32b) can never be satisfied provided μ is reasonably close to λ_l.

Next we consider w to be in a small neighborhood of u_l and take $\mu=\lambda_l$, i.e., $L[w]+\lambda_l\rho w=0$. Let us denote $w=u_l+\delta w$, $\xi=\sigma+\epsilon$ and

$$\delta J(u_l)=\int_\Gamma p\xi u_l^2ds-\int p\sigma u_l^2ds,$$

δw and ϵ being functions of small magnitudes. Then from (31b) we have

$$\delta J(u_l)=\int_\Gamma p\epsilon u_l^2ds\geqq 0.$$

As is evident from the extremal property of $\delta J(u_l)$, it should be possible to express $\delta J(u_l)$ in terms of second-order small quantities. This can be easily shown as follows:

$$\delta J(u_l)=\int_\Gamma pu_l\frac{\partial u_l}{\partial n}ds-\int_\Gamma p\frac{u_l^2}{w}\frac{\partial w}{\partial n}ds$$

$$=\int_\Gamma p\delta w\frac{\partial\delta w}{\partial n}ds-\int_\Gamma p\frac{(\delta w)^2}{w}\frac{\partial w}{\partial n}ds.$$

Then we consider the corresponding variation for the expression (32a)

$$\delta J(w)=\int_\Gamma\frac{p}{\xi}\left(\frac{\partial w}{\partial n}\right)^2ds-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2ds$$

$$=-\int p\delta w\frac{\partial\delta w}{\partial n}ds-\int_\Gamma\frac{p}{\sigma}\left(\frac{\partial\delta w}{\partial n}\right)^2ds,$$

or

$$\delta J(w)=-\int_\Gamma p\frac{\epsilon}{\sigma(\sigma+\epsilon)}\left(\frac{\partial w}{\partial n}\right)^2ds$$

$$=-\int_\Gamma p\epsilon u_l^2\frac{1+[(\partial\delta w/\partial n)/(\partial u_l/\partial n)]^2}{1+\epsilon/\sigma}ds.$$

As long as small variations are considered, $[(\partial\delta w/\partial n)/(\partial u_l/\partial n)]^2\ll 1$ and $\epsilon/\sigma\ll 1$ except in the small neighborhoods of those zeros of $u_l(s)$ where $\partial u_l/\partial n$ is also zero. Let us denote such neighborhoods by η, then

$$\delta J(w)=\delta J_{\Gamma-\eta}+\delta J_\eta$$

where

$$\delta J_\eta = -\int_\eta p\delta w \frac{\partial \delta w}{\partial n} ds - \int_\eta \frac{p}{\sigma}\left(\frac{\partial \delta w}{\partial n}\right)^2 ds$$

and $\delta J_{\Gamma-\eta}$ is a similar expression. As $\partial\delta w/\partial n \to 0$, $\eta \to 0$, $|\delta J_\eta|$ is small in comparison with $|\delta J_{\Gamma-\eta}|$ and $\delta J(w) \to \delta J_{\Gamma-\eta} \to -\int_\Gamma p\epsilon u_l^2 ds$. Since $\int_\Gamma p\epsilon u_l^2 ds \geqq 0$, we must have $\delta J(w) \leqq 0$ at least for small enough variations.

Thus far it has been assumed that ξ is finite wherever σ is finite. We will now remove this restriction.

Let w' be a function which satisfies $L[w'] + \lambda_l \rho w' = 0$ and the given orthogonality conditions. w' is to be in a small neighborhood of u_l; $\partial w'/\partial n = \partial u_l/\partial n + \partial \delta w'/\partial n$, $w' = u_l + \delta w'$. $\partial w'/\partial n = 0$ on Γ wherever $\sigma = 0$. $\xi' = -(1/w')(\partial w'/\partial n)$ may be infinite where σ is finite. By small variations w' may be changed into a function, say w, which satisfies all the conditions imposed on w' and the restriction that $\xi = -(1/w)(\partial w/\partial n)$ is finite where σ is finite. We denote

$$\partial w'/\partial n = \partial w/\partial n + \partial \delta w''/\partial n \qquad (\partial \delta w''/\partial n = \partial \delta w'/\partial n - \partial \delta w/\partial n),$$

$$w' = w + \delta w'' \qquad (\delta w'' = \delta w' - \delta w).$$

As long as $\partial \delta w''/\partial n$ is small, the corresponding zeros of $w(s)$ and $w'(s)$ are near together. Since w may be any suitable function, $\partial \delta w''/\partial n$ may be chosen to be zero on Γ except in those small neighborhoods of the zeros of $w(s)$ and $w'(s)$, say η, where σ and ξ are finite and $(\partial \delta w''/\partial n)$ is to be of suitable form in order that w may satisfy all prescribed conditions.

By simple calculation $\delta J(w')$ is found to be

$$\delta J(w') = \int_\Gamma \frac{p}{\xi}\left(\frac{\partial w'}{\partial n}\right)^2 ds - \int_\Gamma \frac{p}{\sigma}\left(\frac{\partial w'}{\partial n}\right)^2 ds = \delta J(w) + \delta J_\eta',$$

where

$$\delta J_\eta' = -\int_\eta p\delta w'' \frac{\partial \delta w''}{\partial n} ds - \int_\eta \frac{p}{\sigma}\left(\frac{\partial \delta w''}{\partial n}\right)^2 ds + 2\int_\eta \frac{p}{\sigma} \epsilon w \frac{\partial \delta w''}{\partial n} ds.$$

As $\partial \delta w'/\partial n$ and $\partial \delta w/\partial n$ approach zero both $\partial \delta w''/\partial n$ and η approach zero. If $\partial \delta w''/\partial n$ is sufficiently small, η will be so small that $|\delta J_\eta'| < |\delta J(w)|$. Therefore $\delta J(w') \leqq 0$ at least for small enough variations.

Since $\delta J(w')$ can also be expressed in terms of second-order small quantities as

$$\delta J(w') = -\int_\Gamma pw' \frac{\partial w'}{\partial n} ds - \int_\Gamma \frac{p}{\sigma}\left(\frac{\partial w'}{\partial n}\right)^2 ds = -\int_\Gamma p\delta w' \frac{\partial \delta w'}{\partial n} ds - \int_\Gamma \frac{p}{\sigma}\left(\frac{\partial \delta w'}{\partial n}\right)^2 ds$$

and $\delta w'$ is linearly related to $\partial \delta w'/\partial n$, so if $\delta J(w') \leqq 0$ for small $\delta w'$ the same inequality will be true for any large $\Delta w'$ which is a constant multiple of $\delta w'$. Since $\delta w'$ may be any admissible variation, so $\Delta w'$ may be any admissible function. Writing $w = u_l + \Delta w'$ we obtain

$$-\int_\Gamma pw \frac{\partial w}{\partial n} ds - \int_\Gamma \frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2 ds \leqq 0. \tag{33}$$

Here w is no longer restricted to a small neighborhood of u_l and $\xi = -(1/w)(\partial w/\partial n)$ may be left undefined provided that $\partial w/\partial n = 0$ wherever $\sigma = 0$, $L[w] + \lambda_l \rho w = 0$ and w satisfies all those orthogonality conditions imposed on the set of functions u or v.

We then calculate

$$\frac{\partial}{\partial \mu}\int_\Gamma \frac{p}{\xi}\left(\frac{\partial w}{\partial n}\right)^2 ds = -\frac{\partial}{\partial \mu}\int_\Gamma pw \frac{\partial w}{\partial n} ds$$

for fixed $\partial w/\partial n$ with $L[w] + \mu \rho \dot{w} = 0$.

$$-2\frac{\partial}{\partial \mu}\int_\Gamma pw \frac{\partial w}{\partial n} ds = -2\int_\Gamma p \frac{\partial w}{\partial \mu}\frac{\partial w}{\partial n} ds$$

$$= -2D\left[w, \frac{\partial w}{\partial \mu}\right] + 2\mu \int_R\int \rho w \frac{\partial w}{\partial \mu} dxdy$$

$$= -\frac{\partial}{\partial \mu}\left\{ D[w] - \mu \int_R\int \rho w^2 dxdy \right\} - \int_R\int \rho w^2 dxdy$$

$$= -\frac{\partial}{\partial \mu}\int_\Gamma pw \frac{\partial w}{\partial n} ds - \int_R\int \rho w^2 dxdy.$$

Thus

$$-\frac{\partial}{\partial \mu}\int_\Gamma pw \frac{\partial w}{\partial n} ds = -\int_R\int \rho w^2 dxdy,$$

so

$$\frac{\partial}{\partial \mu}\int_\Gamma \frac{p}{\xi}\left(\frac{\partial w}{\partial n}\right)^2 ds < 0. \tag{34}$$

This inequality is valid everywhere except at certain singular points at which the expression

$$\int_\Gamma p \frac{\partial w}{\partial n} ds \int_\Gamma p \frac{\partial w}{\partial n'} G(s, s'; \mu) ds'$$

goes to infinity and changes sign abruptly.

For $\mu = \lambda_l$ the expression

$$-\int_\Gamma pw \frac{\partial w}{\partial n} ds - \int_\Gamma \frac{p}{\sigma}\left(\frac{\partial w}{\partial n}\right)^2 ds$$

is always less than or equal to zero and its value for any fixed $(\partial w/\partial n)$ can only be increased by decreasing μ, assuming μ to be varied over a range which does not contain the singular points. Therefore the condition (32)

can only be satisfied with $\mu \leqq \lambda_l$ no matter what function is chosen for $\partial w/\partial n$ as long as it is an admissible one and μ is less than the lowest singular point lying above λ_l.

We have thus proved our rule for the calculation of a lower bound for the lth eigenvalue. The rule is just the same as that stated in connection with the fundamental mode except that we have, now, $l-1$ additional orthogonality conditions. In passing, it may be noted that we have only used the orthogonality conditions at $\mu=\lambda_l$ in the above proof.

Regarding the question about the perturbation of continuity conditions, we consider the same region R as shown in Fig. 1, and consider the same problem, that is to minimize the expression (14) under the usual appropriate conditions. This is equivalent to a problem of finding a function u which minimizes

$$J[u]=D_I[u]+\int_{\Gamma_I} p\sigma u^2 ds - \int_{\gamma_I} pu(\partial u/\partial n)ds + \text{similar terms for } II \quad (35)$$

under the additional restriction that both u and $\partial u/n$ are continuous on the common boundary γ, i.e.,

$$\int_{\gamma_I+\gamma_{II}} pu(\partial u/\partial n)ds \equiv 0.$$

Here the subscripts I and II refer to regions R_I and R_{II} respectively.

We then perturb the problem such that only the normal derivative of the function should be continuous, not necessarily the function itself and the two surface integrals over γ should add up to zero. Let v denote the new set of trial functions; v certainly contains the set u. Therefore, some functions in v will have

$$J[v] \leqq J[u]_{\min} = J[u_l] = \lambda_l.$$

Let w be a sub-set of functions contained in v such that every function of w has a $J[w]$ which is a minimum among all functions of v which have the same normal derivative $\partial w/\partial n$ on γ. u_l is contained in w.

If the set v has functions with $J[v]<\lambda_l$, w must also have functions with $J[w]<\lambda_l$. This is obvious because there exists a function w corresponding to any such function v such that $\partial w/\partial n=\partial v/\partial n$ on γ and $J[w] \leqq J[v]$.

Since $J[u_l]$ is stationary with respect to arbitrary variations in the normal derivative on γ while other $J[w]$ are not, we must have either $J[w] \leqq J[u_l]$ or $J[w] \geqq J[u_l]$. Otherwise the second variation of $J[w]$ should vanish identically for $w=u_l$, but this can easily be shown to be impossible. Moreover, since it is always possible to find a function w which satisfies $J[w] \leqq J[u_l]$, e.g., $\partial w/\partial n=0$ on γ is such a function, therefore we must have

$$J[w] \leqq J[u_l].$$

The case where $J[w]$ is close to a higher eigenvalue is to be excluded.

We denote $J[w]=\mu$. It follows directly from the usual variational treatment that we must have

$$L[w]+\mu\rho w=0 \text{ in } R, \quad \partial w/\partial n+\sigma w=0 \text{ on } \Gamma.$$

Under these conditions

$$w_I = \int_\gamma p \frac{\partial w_I}{\partial n'} G_I(s, s'; \mu)ds',$$

$$w_{II} = \int_\gamma p \frac{\partial w_{II}}{\partial n'} G_{II}(s, s'; \mu)ds',$$

$$\left(\frac{\partial w_I}{\partial n} = -\frac{\partial w_{II}}{\partial n} \text{ on } \gamma\right),$$

where $G_I(x, y; x', y'; \mu)$ and $G_{II}(x, y; x', y'; \mu)$ are Green's functions of the same equation $L[u]+\mu\rho u=0$, such that G satisfies the appropriate boundary conditions on Γ and $\partial G_I/\partial n=\partial G_{II}/\partial n=0$ on γ. Both G_I and G_{II} exist because we have postulated that the problem can be solved exactly for separate regions once we impose boundary conditions on γ.

Hence we may conclude that the condition

$$0=\int_{\gamma_I+\gamma_{II}} pw\frac{\partial w}{\partial n}ds = \int_\gamma p\frac{\partial w_I}{\partial n}ds \int_\gamma p\frac{\partial w_I}{\partial n'}G_I(s, s'; \mu)ds' + \int_\gamma p\frac{\partial w_{II}}{\partial n}ds \int_\gamma p\frac{\partial w_{II}}{\partial n'}G_{II}(s, s'; \mu)ds' \quad (36)$$

can never be satisfied with $\mu>\lambda_l$ for any admissible trial function used for $\partial w/\partial n$ on γ, provided μ is reasonably close to λ_l.

We may further point out that the two processes of perturbation, that of perturbing boundary conditions and of perturbing continuity conditions, may be carried out together for the same problem and the above proofs can easily be extended to such cases.

IV. SERIES EXPANSION AND ORTHOGONALITY CONDITIONS

We will show in this section how we can avoid the difficulty of having no exact knowledge about the lower eigenfunctions. We consider the case of perturbation of continuity conditions, the other case, namely, the perturbation of boundary conditions is somewhat simpler. The variational expression for the upper-bound case has properties closely connected to those of the expression for calculating the lower bound, so we will discuss them together.

We consider the lth mode for the same problem as before. Either $w(s)$ or $\partial w/\partial n$ is to be continuous on γ (Fig. 1) and $\partial w/\partial n+\sigma w(s)=0$ on Γ. In either case, we

may expand $w_I(s)$ and $w_{II}(s)$ into series in terms of different sets of orthogonal functions. This implies that we may expand G_I, G_{II} or $\partial G_I/\partial n$, $\partial G_{II}/\partial n$ in terms of their respective sets of eigenfunctions.

Thus we write

$$\begin{aligned} w_I(s) &= \sum_{j=1}^{\infty} b_{I,j}\phi_{I,j}(s) \text{ on } \gamma \\ w_{II}(s) &= \sum_{j=1}^{\infty} b_{II,j}\phi_{II,j}(s) \text{ on } \gamma, \end{aligned} \tag{37}$$

where b_j's are functions of λ but not of position, and $\phi_j(s)$ are the orthonormal coordinate functions such that

$$\int p\phi_{I,i}\phi_{I,j}ds = \delta_{ij}$$

and

$$\int_{\gamma} p\phi_{II,i}\phi_{II,j}ds = \delta_{ij}. \tag{38}$$

The existence of such functions depends, of course, on the assumption that exact solutions for the modified problem are known for R_I and R_{II} separately. In general, $\phi_{I,i}$ are not orthogonal to $\phi_{II,i}$.

We denote as before the first $(l-1)$ eigenfunctions of the original problem by u_i $(i=1, 2, \cdots, l-1)$. $u_i(s)$ may also be expanded in terms of ϕ_I's or ϕ_{II}'s. Thus

$$\begin{aligned} u_i(s) &= \sum_{j=1}^{\infty} \epsilon_{I,ij}\phi_{I,j}(s) \text{ on } \gamma \\ &= \sum_{j=1}^{\infty} \epsilon_{II,ij}\phi_{II,j}(s) \text{ on } \gamma \\ & i=1, 2, \cdots, l-1. \end{aligned} \tag{39}$$

Instead of (37) $w(s)$ may be expressed in a somewhat different form by using the $(l-1)$ functions $u_i(s)$ to replace any $(l-1)$ coordinate functions ϕ_j. Let $l-1=2v$ or $2v-1$ according as l is odd or even. We replace the first $(l-1)\phi_j$'s by u_i's and write $w(s)$ as

$$\begin{aligned} w_I(s) &= \sum_{i=1}^{l-1} c_i u_i(s) + \sum_{j=v+1}^{\infty} c_{I,j}\phi_{I,j}(s) \\ w_{II}(s) &= \sum_{i=1}^{l-1} c_i u_i(s) + \sum_{j=v \text{ or } v+1}^{\infty} c_{II,j}\phi_{II,j}(s). \end{aligned} \tag{40}$$

Substituting the expression (39) for $u_i(s)$ into (40) and comparing with (37), we obtain

$$\begin{aligned} b_{I,j} &= \sum_{i=1}^{l-1} c_i\epsilon_{I,ij}, \quad j=1, 2, \cdots, v; \\ b_{II,j} &= \sum_{i=1}^{l-1} c_i\epsilon_{II,ij}, \quad j=1, 2, \cdots, v-1 \text{ or } v. \end{aligned} \tag{41}$$

If the determinant $|\epsilon_{ij}| \neq 0$, the above $l-1$ equations will just be enough for the determination of $l-1$ c_i's. In case all b_j's vanish, c_i's can only be zero. If $|\epsilon_{ij}| = 0$, we may write $w(s)$ as

$$\begin{aligned} w_I(s) &= \sum_{i=1}^{l-1} c_i u_i(s) + c_N u_N(s) + \sum_{j=v+1}^{\infty} c_{I,j}\phi_{I,j}(s), \\ w_{II}(s) &= \sum_{i=1}^{l-1} c_i u_i(s) + c_N u_N(s) + \sum_{j=v \text{ or } v+1}^{\infty} c_{II,j}\phi_{II,j}(s), \end{aligned}$$

with N being any integer $\geqq l+1$, such that the rank of the matrix $(\epsilon_{ij}, \epsilon_{N,j})$ is $l-1$. Then we have in place of (41)

$$\begin{aligned} b_{I,j} &= \sum^{l-1} c_i\epsilon_{I,ij} + c_N\epsilon_{I,Nj}, \quad j=1, 2, \cdots, v, \\ b_{II,j} &= \sum_{i=1}^{l-1} c_i\epsilon_{II,ij} + c_N\epsilon_{II,Nj}, \quad j=1, 2, \cdots, v-1 \text{ or } v. \end{aligned}$$

By assigning a certain value, zero or non-zero, to one suitable coefficient c_i we will be able to determine the other c_i's and c_N uniquely. The rest of the discussion will be seen to be essentially the same for the two different cases. Therefore, we may assume without loss of generality that $|\epsilon_{ij}| \neq 0$ and $w(s)$ can be expressed in the form of (40).

Similarly, we can also express $\partial w/\partial n$ in terms of ϕ_j's and $u_i(s)$ as

$$\begin{aligned} \frac{\partial w_I}{\partial n} &= -\sum_{i=1}^{l-1} c_i\chi_{I,i}u_i(s) - \sum_{j=v+1}^{\infty} c_{I,j}\sigma_{I,j}\phi_{I,j}(s), \\ \frac{\partial w_{II}}{\partial n} &= -\sum_{i=1}^{l-1} c_i\chi_{II,i}u_i(s) - \sum_{j=v \text{ or } v+1}^{\infty} c_{II,j}\sigma_{II,j}\phi_{II,j}(s), \end{aligned} \tag{42}$$

where σ_j are functions of λ but not of position, just as are c_i and c_j. Since $u_i(x, y)$ are eigenfunctions, $\partial u_i/\partial n$ must be continuous on γ, so $\chi_{I,i} = -\chi_{II,i}$.

With such expansions for $w(s)$, $\partial w/\partial n$ and $u_i(s)$ at hand, we proceed to find what conditions $w(s)$ must be subject to in order that the orthogonality conditions may be satisfied.

By Green's theorem, we have

$$\begin{aligned} &\int_{\Gamma_{I+\gamma}} p\left(w\frac{\partial u_i}{\partial n} - u_i\frac{\partial w}{\partial n}\right)ds \\ &= \iint_{R_I} \{wL[u_i] - u_iL[w]\}dxdy, \quad (i=1, 2, \cdots, l-1) \end{aligned}$$

and similar equations for R_{II}. Combining the equations for I and II together and putting $L[u_i]+\lambda_i\rho u_i=0$ and $L[w]+\lambda\rho w=0$, we obtain

$$\int_{I+II} p\left(w\frac{\partial u_i}{\partial n} - u_i\frac{\partial w}{\partial n}\right)ds = -(\lambda_i-\lambda)\iint_R \rho w u_i dxdy.$$

The latter equation is further reduced to

$$\int_{\gamma_I+\gamma_{II}} p\left(w\frac{\partial u_i}{\partial n}-u_i\frac{\partial w}{\partial n}\right)ds=-(\lambda_i-\lambda)\int_R\int \rho w u_i dxdy,$$

because w is to satisfy the appropriate conditions, i.e., $\partial w/\partial n+\sigma w=0$ on Γ. So we must have

$$\int_{\gamma_I+\gamma_{II}} p\left(w\frac{\partial u_i}{\partial n}-u_i\frac{\partial w}{\partial n}\right)ds=0$$

in order that w will be orthogonal to u_i. Furthermore, since either w or $\partial w/\partial n$ is continuous on γ, we must have both

$$\int_{\gamma_I+\gamma_{II}} pw\frac{\partial u_i}{\partial n}ds=0 \quad \text{and} \quad \int_{\gamma_I+\gamma_{II}} pu_i\frac{\partial w}{\partial n}ds=0. \tag{43}$$

Substituting (40) and (42) into (43), we obtain

$$\int_{\gamma_I+\gamma_{II}} pw\frac{\partial u_i}{\partial n}ds=-\int_\gamma p\chi_{I,i}u_i(s)\times\{\sum_{j=\nu+1}^{\infty} c_{I,j}\phi_{I,j}(s)\}ds+\text{similar term for } II=0 \tag{44}$$

and

$$\int_{\gamma_I+\gamma_{II}} pu_i\frac{\partial w}{\partial n}ds=-\int_\gamma pu_i(s)\times\{\sum_{j=\nu+1}^{\infty} c_{I,j}\sigma_{I,j}\phi_{I,j}(s)\}ds+\text{similar term for } II=0. \tag{45}$$

Now we calculate

$$\begin{aligned}J(\lambda)&=\int_{\gamma_I+\gamma_{II}} pw\frac{\partial w}{\partial n}ds\\&=-\sum_{i=1}^{l-1}c_i\int_\gamma pu_i(s)\{\sum_{j=1}^{l-1}c_j\chi_{I,j}u_j(s)\}ds-\sum_{j=\nu+1}^{\infty}(c_{I,j})^2\sigma_{I,j}\\&\quad-\sum_{j=\nu+1}^{\infty}c_{I,j}\sigma_{I,j}\int_\gamma p\{\sum_{i=1}^{l-1}c_iu_i(s)\}\phi_{I,j}(s)ds\\&\quad-\sum_{j=\nu+1}^{\infty}c_{I,j}\int_\gamma p\{\sum_{i=1}^{l-1}c_i\chi_{I,i}u_i(s)\}\phi_{I,j}(s)ds\\&\quad+\text{similar terms for } II.\end{aligned}$$

By virtue of (44) and (45) and $\chi_{I,j}=-\chi_{II,j}$, the above equation is reduced to

$$J(\lambda)=-\sum_{j=\nu+1}^{\infty}(c_{I,j})^2\sigma_{I,j}-\sum_{j=\nu \text{ or } \nu+1}^{\infty}(c_{II,j})^2\sigma_{II,j}. \tag{46}$$

Thus, if w is orthogonal to $u_i(i=1, 2, \cdots, l-1)$, the series expansion for $J(\lambda)$ will contain none of the first $(l-1)$ terms. The converse statement, however, is in general not true.

Let us rewrite (40) as

$$\begin{aligned}w_I(s)&=\sum_1^{l-1}c_iu_i(s)+f_I(s)\\w_{II}(s)&=\sum_1^{l-1}c_iu_i(s)+f_{II}(s)\end{aligned} \tag{40'}$$

and (42) as

$$\begin{aligned}\frac{\partial w_I}{\partial n}&=-\sum_1^{l-1}c_i\chi_{I,i}u_i(s)+\frac{\partial f_I}{\partial n}\\\frac{\partial w_{II}}{\partial n}&=-\sum_1^{l-1}c_i\chi_{II,i}u_i(s)+\frac{\partial f_{II}}{\partial n}.\end{aligned} \tag{42'}$$

If $f(s)$ or $\partial f/\partial n$ happens to be of suitable form, w will be orthogonal to $u_i(i=1, 2, \cdots, l-1)$; otherwise not. The function $w(s)$, which approximates $u_l(s)$, can be exactly right only when $J(\lambda)$ has no first $(l-1)$ terms.[16] Thus if we use $f(s)$ in place of $w(s)$ as the trial function, we will be able to get to the exact form of $J(\lambda)$, i.e.,

$$\int_{\gamma_I+\gamma_{II}} pu_l\frac{\partial u_l}{\partial n}ds\equiv 0,$$

by varying $f(s)$. It is to be noted that $f(s)$ does not approximate $u_l(s)$ but the function

$$u_l(s)-\sum^{l-1}c_iu_i(s).$$

Let us rewrite the expression for $J(\lambda)$ also. If we write $J(\lambda)$ for the upper-bound case, where w is matched

[16] For the case of perturbing boundary conditions, the equation $J(\lambda)=-\int_\Gamma p\sigma w^2ds$ or $J(\lambda)=-\int_\Gamma p/\sigma(\partial w/\partial n)^2ds$ will be reduced to

$$-\sum_{j=l}^{\infty}\sigma_jc_j^2=-\int_{\Gamma'}p\sigma(\sum_{j=l}^{\infty}c_j\phi_j)^2ds$$

or

$$-\sum_{j=l}^{\infty}\sigma_jc_j^2=-\int_{\Gamma'}p/\sigma(\sum_{j=l}^{\infty}c_j\sigma_j\phi_j)^2ds,$$

Γ' being that part of Γ where the boundary condition is perturbed and a complete set of appropriate orthogonal functions ϕ_j exist. Thus

$$J(\lambda)+\int_\Gamma p\sigma w^2ds=\int_{\Gamma'}p\sigma f^2ds-\sum_{j=l}^{\infty}\sigma_j\left(\int_{\Gamma'}pf\phi_jds\right)^2$$

or

$$J(\lambda)+\int_\Gamma p/\sigma(\partial w/\partial n)^2ds=\int_{\Gamma'}p/\sigma(\partial f/\partial n)^2ds-\sum_{j=l}^{\infty}1/\sigma_j\left(\int_{\Gamma'}p(\partial f/\partial n)\phi_jds\right)^2$$

will have no first $(l-1)$ terms if w is orthogonal to u_i $(i=1, 2, \cdots, l-1)$. The rest of the discussion will be similar to that for the case of perturbing continuity conditions.

on γ, as

$$J_+(\lambda)=-\sum_1^\infty \sigma_{I,j}(\lambda)\left(\int_\gamma pw\phi_{I,j}ds\right)^2 -\sum_1^\infty \sigma_{II,j}(\lambda)\left(\int_\gamma pw\phi_{II,j}ds\right)^2, \quad (47)$$

the corresponding expression for the lower-bound case, where $\partial w/\partial n$ is matched, will be

$$J_-(\lambda)=-\sum_1^\infty \frac{1}{\sigma_{I,j}(\lambda)}\left(\int_\gamma p\cdot\frac{\partial w}{\partial n}\cdot\phi_{I,j}ds\right)^2 -\sum_1^\infty \frac{1}{\sigma_{II,j}(\lambda)}\left(\int_\gamma p\cdot\frac{\partial w}{\partial n}\cdot\phi_{II,j}ds\right)^2. \quad (48)$$

It is obvious that the corresponding terms of $J_+(\lambda)$ and $J_-(\lambda)$ always have the same sign. For the fundamental mode there is no question about the orthogonality conditions, so only the first term can have a sign different from all the rest. Otherwise we would get at least two values for λ. This could not be possible if w has not zeros inside R. As λ increases, the higher terms will change sign. The second term will have changed sign for the second mode and so the lth term changed for the lth mode. As long as the first $l-1$ terms in $J(\lambda)$ are missing, there can be only one term which has different sign from all the rest. As shown by Schwinger such forms of $J(\lambda)$ are indeed definite for a certain fixed value of λ, being λ_l in the present case. Thus, if we can show that $\partial J_+(\lambda)/\partial\lambda$ has definite sign and the sign is always opposite to that of $\partial J_-(\lambda)/\partial\lambda$, then the value of λ calculated by solving $J_+=0$ or $J_-=0$ will always be greater or lower than the true value.

Though the above result may be anticipated from the proof we had for (34), it is desirable to determine how each term of the series varies with λ. We consider a single coefficient $\sigma_j(\lambda)$ used in (47) and calculate $\partial/\partial\lambda\sigma_j(\lambda)$. Since σ_j satisfies $\partial\phi_j/\partial n+\sigma_j\phi_j=0$ on γ by definition, so

$$\int_\gamma p\phi_j\frac{\partial\phi_j}{\partial n}ds=-\int_\gamma p\sigma_j\phi^2 ds.$$

Since $\phi_j(x, y)$ satisfies the differential equation, we have by Green's theorem

$$\iint_{I\text{ or }II}\{p(\nabla\phi_j)^2+q\phi_j^2\}dxdy +\iint_{I\text{ or }II}\phi_j L[\phi_j]dxdy=\int_{I\text{ or }II}p\phi_j\frac{\partial\phi_j}{\partial n}ds$$

i.e.,

$$D_{I\text{ or }II}[\phi_j,\phi_j]-\lambda\iint_{I\text{ or }II}\rho\phi_j^2dxdy=\int_{I\text{ or }II}p\phi\frac{\partial\phi_j}{j\partial n}ds.$$

Differentiating with respect to λ, we obtain

$$\frac{\partial}{\partial\lambda}D_{I\text{ or }II}[\phi_j,\phi_j]-\iint_{I\text{ or }II}\rho\phi_j^2dxdy -2\lambda\iint_{I\text{ or }II}\rho\phi_j\frac{\partial\phi_j}{\partial\lambda}dxdy=\frac{\partial}{\partial\lambda}\int_{I\text{ or }II}p\phi_j\frac{\partial\phi_j}{\partial n}ds.$$

Since $\phi_j(x, y)$ as $w(x, y)$ satisfies the appropriate boundary conditions on Γ, i.e., $\partial\phi_j/\partial n+\sigma\phi_j=0$ on Γ_I and Γ_{II}, the right-hand side of the above equation becomes

$$\frac{\partial}{\partial\lambda}\int_{I\text{ or }II}p\phi_j\frac{\partial\phi_j}{\partial n}ds =-\frac{\partial}{\partial\lambda}\int_{\Gamma_I\text{ or }\Gamma_{II}}p\sigma\phi_j^2ds-\frac{\partial}{\partial\lambda}\int_\gamma p\sigma_j\phi_j^2ds$$

while the left-hand side is

$$2D_{I\text{ or }II}\left[\phi_j,\frac{\partial\phi_j}{\partial\lambda}\right]-\iint_{I\text{ or }II}\rho\phi_j^2dxdy-2\lambda\iint_{I\text{ or }II}\rho\phi_j\frac{\partial\phi_j}{\partial\lambda}dxdy$$

$$=2\int_{I\text{ or }II}p\frac{\partial\phi_j}{\partial\lambda}\cdot\frac{\partial\phi_j}{\partial n}ds-\iint_{I\text{ or }II}\rho\phi_j^2dxdy$$

$$=-\frac{\partial}{\partial\lambda}\int_{\Gamma_I\text{ or }\Gamma_{II}}p\sigma\phi_j^2ds-\frac{\partial}{\partial\lambda}\int_\gamma p\sigma_j\phi_j^2ds +\frac{\partial\sigma_j}{\partial\lambda}\int_\gamma p\phi_j^2ds-\iint_{I\text{ or }II}\rho\phi_j^2dxdy.$$

In the last step of derivation, we have made use of the fact that σ is the given boundary condition, and so independent of λ, while σ_j depends on λ but not on position.

Therefore by equating the two sides we obtain

$$\frac{\partial\sigma_j}{\partial\lambda}\int_\gamma p\phi_j^2ds=\iint_{I\text{ or }II}\rho\phi_j^2dxdy.$$

Since $p>0$ and $\rho>0$, so

$$\partial\sigma_j/\partial\lambda>0,$$

so also

$$(\partial/\partial\lambda)(1/\sigma_j)<0.$$

Thus we obtain

$$\partial J_+(\lambda)/\partial\lambda<0 \quad (49)$$

while

$$[\partial J_-(\lambda)]/\partial\lambda>0, \quad (50)$$

because every term of J_+ or J_- changes in the same sense.

Now we have proved that, by dropping out the first $l-1$ terms in J_+ or J_- the error in the calculated value of λ will always have one sign. In fact we can be spared the above proof by utilizing the maximum-minimum property of eigenvalues. According to this property the sense of error for the lower-bound calculation cannot be changed by using a trial function which is not exactly orthogonal to the $(l-1)$ u_i's. Since J_- is positive definite, $\partial J_-/\partial\lambda$ must be positive also. By the reciprocal nature of the role played by σ_j in J_+ and J_- we must have $\partial J_+/\partial\lambda<0$. Thus we obtain the same result in a more elegant way.

The original requirements, that $w(x, y)$ should be orthogonal to all the lower eigenfunctions, have been reduced to the requirements that $f(s)$, which approximates

$$u_l(s)-\sum_1^{l-1} c_i u_i(s),$$

should be orthogonal to the first $l-1$ known functions ϕ_j's. This not only simplifies the process but also makes the upper-bound calculation possible without knowing the true eigenfunctions.

Using such a trial function we are assured that the calculated value of λ will always lie above or always lie below λ_l but not always very close. In any case we must use good enough trial functions. The higher terms of the series will be determined by the singular character of the function while the lower ones by the orthogonality requirements. Thus $f(s)$ (and so $w(s)$) can actually be chosen to be sufficiently near to the true function.

We may further improve the accuracy by successive approximations. For example, we take

$$f'(s)=f(s)+\beta\sum_{v+1}^{2v+1 \text{ or } 2v+2}(c_{I,j}\phi_{I,j}(s)),$$

where

$$\sum_{v+1}^{2v+1 \text{ or } 2v+2} c_{I,j}\phi_{I,j}(s)$$

is to be orthogonal to $\phi_{II,i}(i=1, 2, \cdots, v \text{ or } v+1)$, so $c_{I,j}(j=v+1, v+2, \cdots, 2v+1 \text{ or } 2v+2)$ are known to a constant factor. Apply variation on $J(\lambda)$ with respect to β, i.e., put $\delta_\beta J(\lambda)=0$. This determines β and so all the coefficients $\beta c_{I,j}$. The value calculated by using $f'(s)$ will be more accurate than by using $f(s)$. This process can be carried on successively to whatever accuracy we may desire. In most practical cases, however, the trial function may be chosen to be sufficiently accurate, so there is seldom need to go further than a second approximation.

The author wishes to express his deep gratitude to the late Professor W. W. Hansen, from whom he received invaluable advice and friendly encouragement.

Letters to the Editor

Single Crystal X-Ray Techniques for Use with Alloys Requiring Protective Atmospheres*

JOSEPH SINGER AND F. H. ELLINGER
Los Alamos Scientific Laboratory, University of California, Los Alamos, New Mexico
January 27, 1950

PROCEDURES similar to those to be described undoubtedly have been used by others; it was felt, however, that their accessibility in the literature might be desirable.

Our procedures are complicated by the need for handling our alloys in protective atmospheres. An ordinary type of "dry box" equipped with rubber gloves is the preparation chamber through which the inert gas constantly flows. Single crystal fragments of intermetallic compounds have been obtained by shattering ordinary ingots with a hammer. The fragments are viewed in a glass dish against a background of dull black paper. With the aid of suitable tweezers a glass fiber is touched to thin shellac and brought up to a selected fragment. With ingots of small grain size, the fragment selected should be as small as can be seen, preferably with the aid of magnification. If now the mounted fragment were to be dropped into a capillary tube of the type shown in Fig. 1, in which the evacuation and sealing-off is to be done, the specimen would very likely be lost. The following are two ways of controlling the placing of the specimen inside the capillary tube.

One procedure is to place the mounting fiber into a shorter, thin-walled plastic tube which has been formed around a wire smaller than the glass capillary. The metallic fragment is set barely to protrude beyond the plastic and the back end of its mount is fastened to the plastic by clamping. This assemblage can be dropped safely into the glass tube. Figure 1 illustrates this procedure.

A second method consists of using a long fiber, drawn fine only at the end where the specimen is to be attached. Each such fiber is made the right length for a particular capillary tube so that, when the specimen is at a desired position, the back end of the fiber is at the mouth of the tube. With the specimen suitably situated, the tube is held horizontally and shellac brushed on the back end where the fiber touches the wall at the entrance to the tube. When the shellac is dry, the assembly can be held vertically for the evacuation and sealing-off without displacing the specimen. The sealing-off will fasten the fiber within the capillary so that the sample is held rigidly at a desired position.

It is frequently necessary to mount a specimen about a second axis, once a set of single crystal data has been obtained for one

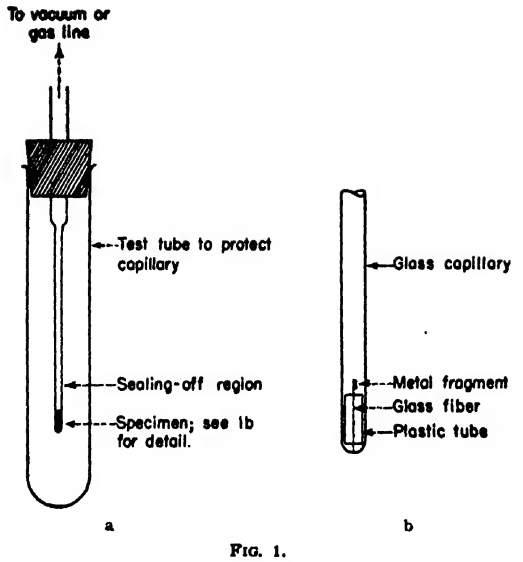


FIG. 1.

axis. With fragments of alloys the usual crystallographic aids for locating an axial direction are not available. However, the position of an axis can be deduced from the previous data. This direction can be set parallel to the x-ray beam. A long, rather heavy glass fiber is set up on a lump of putty-like material and worked over towards the crystal mount until the two pieces of glass touch and the long fiber parallels the selected direction. A small quantity of shellac is deposited around the intersection and is allowed to harden thoroughly. Finally the old mount is carefully detached from the goniometer head, the new fiber shortened, and the crystal aligned on its new axis.

With a little practice one can perform the dry box operations skillfully despite the intervening rubber gloves. The effort is amply repaid with the type of x-ray diffraction data resulting.

* This letter is based on work performed at Los Alamos Scientific Laboratory of the University of California under contract W-7405-Eng-36.

Errata: On the Diffraction of an Electromagnetic Wave through a Plane Screen

[J. App. Phys. 20, 760 (1949)]

J. W. MILES
Department of Engineering, University of California, Los Angeles, California

THE scattering cross section for plane, electromagnetic waves normally incident upon a circular aperture in an infinite, plane, perfectly conducting screen given by us in a recent paper[1] was calculated erroneously. This error, in turn, was due to an error

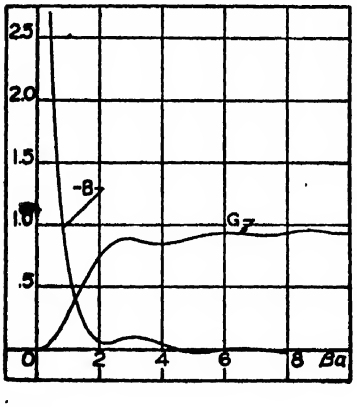


FIG. 9. The conductance and susceptance of a circular aperture for normal incidence of a plane, electromagnetic wave.

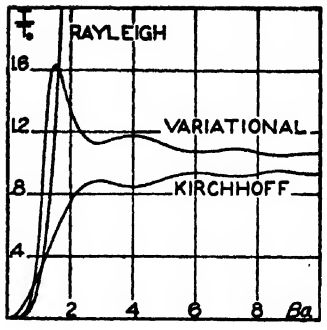


FIG. 10. Scattering cross section (relative to geometric) of a circular aperture for normal incidence of a plane, electromagnetic wave.

in the computation of the aperture susceptance (B). The results have been recalculated, and Figs. 9 and 10 in the original paper should be replaced by those given here.

The error in our calculations was brought to our attention by Professor Harold Levine (Harvard University), who has carried out similar (variational) calculations,[2] using as first and second approximations to the aperture field $[1-(r/a)^2]^{\frac{1}{2}}$ and $A_1[1-(r/a)^2]^{\frac{1}{2}}+A_2[1-(r/a)^2]^{\frac{3}{2}}$, respectively. Our calculation, which uses the static approximation $[1-(r/a)^2]^{\frac{1}{2}}$ for the leading term for small βa and the incident field for the remaining terms, does not predict the first maximum of the scattering cross section as accurately as that of Levine. (We obtain $(T/T_0)_{max}=1.63$ at $\beta a=1.5$; Levine obtains $(T/T_0)_{max}=1.78$ at $\beta a=1.4$.) But it is somewhat more accurate (relative to Levine's first approximation) for large βa, all of which is as should be expected from the nature of the field approximations.

Two other errors in the paper should be pointed out. Equation (86b) should have read

$$B_e(\beta a)=\sum_{m=0}^{\infty}\frac{(-)^m(\beta a)^{2m+1}}{(2m+2)[\Gamma(m+\frac{3}{2})]^2}-(3\pi/8\beta a). \qquad (86b)$$

(This error was typographical and not connected with the calculations.) Moreover, contrary to our statement, Magnus[3] *did* solve the integral equation which he formulated for the Sommerfeld problem,[3] and we should have stated merely that his solution was not as direct as the Wiener-Hopf approach.

[1] J. App. Phys. 20, 760–771 (1949).
[2] Progress Report 6, Cruft Laboratory, Harvard University, Contract N5-ORI-76 (January 1, 1948).
[3] W. Magnus, Zeits. f. Physik 117, 168 (1941). See also W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Chelsea Publishing Company, New York, 1949), pp. 142, 143.

Erratum: Distribution of Phases in Two-Phase Solids

[J. App. Phys. 20, 1206–1208 (1949)]

LEONARD D. JAFFE
Watertown Arsenal Laboratory, Watertown, Massachusetts

AN error in proof-reading made the inequality on page 1208, line 15, inconsistent with the rest of the paper. It should read $r_1<7\times10^{-6}$ cm.

New Books

Letter Symbols for Aeronautical Sciences

BY SEVEN SPONSORING OR COLLABORATING ORGANIZATIONS. Pp. 20. Published by American Standards Association, New York, 1950. Price $1.25.

Journal
of
Applied Physics

Volume 21, Number 6 June, 1950

Special Issue on High Polymer Physics

The following thirteen papers were presented at the meeting of the Division of High Polymer Physics of the American Physical Society in New York City, February 2–4, 1950.

Random Noise in Dielectric Materials

R. F. Boyer
The Dow Chemical Company, Midland, Michigan

When thin films of polar plastic materials are subjected to d.c. potentials of 10 to 300 volts per mil they exhibit current fluctuations (with frequencies in the range of 60 to 1000 c.p.s.) which last for minutes and which even persist for a few seconds on removing the voltage. These fluctuations are about 1000 fold above the noise level of the circuit and are about 1 percent of the steady d.c. current. They eventually die out but can usually be restored by reversing the polarity. This noise is greater the more polar the nature of the polymer and the higher its moisture content. The effect seems to arise from water or impurity ions which can jump whenever a hole opens between polymer chains and is thus an activated rate process sufficiently slowed down by the highly viscous nature of the polymeric medium to be readily observable. On an oscillograph screen it is possible to see both the forward and the backward diffusion. This noise is most pronounced in Cellophane, Nylon, and copolymers of vinylidene chloride with acrylonitrile, but also appears with polystyrene stored at 86 percent relative humidity. Films of methylcellulose do not show the effect. Numerically the effect seems to arise from the random motion of groups of several hundred ions as a unit. As an extension of these ideas it is suggested and demonstrated that the moisture vapor diffusion constants for a series of polymers increase with the frequency of the loss factor maximum while the moisture vapor permeability increases with the direct current conductivity.

INTRODUCTION

WHEN a direct current potential of the order of 100 volts/mil is applied to a thin film of dielectric material, as in making a resistivity test, it is always observed that there is an initial high current which drops off rapidly to an approximately steady value. Thus it is customary to measure the current after one minute duration and use this value in computing the specific resistivity. We find however that this "steady" current actually fluctuates in a seemingly random manner. The magnitude of these fluctuations is usually about one percent or less of the average current flow. The duration of this random current is a matter of minutes and in some cases up to an hour or more: depending on the kind of material, moisture content, thickness of sample, and applied voltage.

It is the purpose of this paper to discuss in a general way the main features of this random noise effect; to relate it with the general mechanism of moisture vapor transmission through polymers; and to show how the moisture vapor transmission is numerically related to the d.c. conductivity and the a.c. loss factor of the polymer. Since this random noise effect is apparently a new one, there are many aspects of it which are still obscure to us.

HISTORY

In 1945 R. H. Campbell of the Webster Electric Company informed us that samples of Saran film (vinylidene chloride—10 percent vinyl chloride copolymer) used in condenser A of the high gain circuit of Fig. 1 caused a plopping noise in the loudspeaker whereas polystyrene did not show this effect. Mr. Campbell suspected moisture. We considered it a possible dipole analog of the Barkhausen effect wherein whole regions of dipoles would rotate as a unit. Indeed we tried to look for such effects on a 1000-cycle bridge circuit which used a cathode ray oscillograph as a null indicator, but we were unsuccessful. Two years ago it was decided to repeat Mr. Campbell's observations by using an identical circuit except that the input resistance of our amplifier was 5 meg. We checked his findings. Moreover, we also connected the amplifier output to a Brush recording

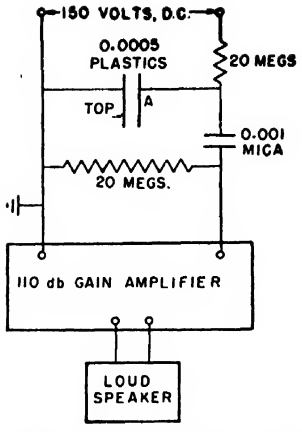


FIG. 1. Circuit used for noise studies. A is the experimental condenser. The amplifier has a 5-megohm input resistance, and cuts off below 30 cycles. In later studies the 0.001 mica blocking condenser was eliminated and the experimental condenser substituted in its place to simplify the circuit.

oscillograph[1] and obtained a permanent record of these plops. They were relatively weak and only a factor of two above the basic noise of the circuit which was estimated to be about a microvolt.

We next tried some other types of vinylidene chloride copolymers, particularly the ones with high acrylonitrile content[2] which would have a high dipole moment. These proved to be quite noisy, with voltage fluctuations of several millivolts. A number of other plastic materials such as Nylon, rubber hydrochloride, paper, Cellophane, plasticized polyvinyl chloride, etc., were then checked. Each of these proved to be more noisy than the original saran sample, although each had a characteristic noise. In fact, it could be said that both the "noise" in the loudspeaker and the pattern drawn by Brush recorders were characteristic of a material. For example, Saran gave an occasional dull plop, Nylon a hissing or frying noise, polyvinyl chloride a scratchy noise. We realize the inadequacy of such verbal descriptions but an examination of typical records may clarify these descriptions. There was, as will be seen later in Fig. 6, a difference in the tracings made by the 60- and 120-cycle pen arms of the Brush recorder. Obviously each "noise" has a frequency distribution spectrum: the Brush oscillograph measures predominantly the low frequency component (say up to 100 cycles); the loudspeaker responds well up to 10.000 cycles, etc. In some instances visual observation of the screen of a fast cathode-ray oscillograph was employed to study the high frequency components in more detail.

Figure 2 shows a time sequence on a sample of coated Cellophane 1.5 mils thick to which 150 volts was applied. Each horizontal line represents one foot of record or one minute of elapsed time while the numbers at the end of the line represent the number of feet of record which has been deleted. The symbols along the chart indicate various operations: B-150 signifies the application of 150 volts d.c. with the bottom electrode negative. The vertical scale is voltage, the maximum peaks representing about 25 mv. At B^+ the polarity on the sample was reversed, leading to considerable noise. Reversal of the voltage as in line 3 leads to a gradual diminution of the noise; the reversal to B^+ in line 6 restores the noise to full intensity with little inclination to die off. Reversal in line 7 again leads to a decaying noise.

The various details of Fig. 2 are characteristic of the behavior of most of the materials, the variation from material to material lying in the direction of intensity of the noise, duration under a given applied voltage, and what we might call the general random behavior of the noise. The main features of Fig. 2, generalized to include most of our experience, are as follows.

1. The noise is usually small at the beginning and its onset may require either a fixed voltage applied for some seconds, or several reversals of the voltage, or increasing the applied voltage to a high value—say 300 volts until the noise starts, and then reducing the voltage.

2. There are usually directional effects in the films. This is quite understandable in a coated Cellophane whereas with a material such as Nylon it might be presumed that manufacturing operations give a different structure to the two surfaces.

3. The noise usually dies out with time but can be restored in general by reversal of the polarity or by increasing the applied voltage. Eventually however, most samples become dead and nothing will restore the

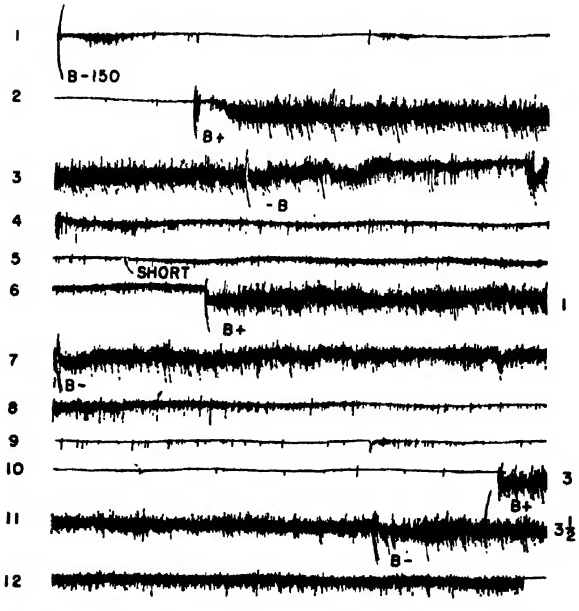


FIG. 2. Noise pattern obtained on a sample of coated Cellophane whose uncoated side was on the bottom electrode. Each horizontal traverse represents one minute of time (one foot of chart paper), the numbers at the right representing feet of record deleted. The ordinates are voltage, the maximum spread being about 25 mv. The symbols B+ and B− refer to the polarity of the bottom electrode.

[1] The Brush Development Company, Cleveland, Ohio.
[2] United States Patent 2,238,020.

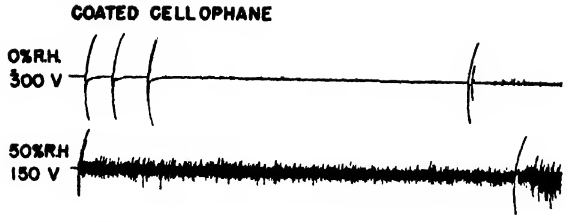


FIG. 3. Effect of relative humidity at which sample was stored on noise pattern of coated Cellophane.

noise. This abrupt termination of the noise is frequently preceded by a particularly loud burst of noise, and one gets the impression of actual dielectric breakdown although we have never noted any visual signs of breakdown.

4. One further effect was noted especially on polyethylene at 86 percent R.H. and 300 volts: the noise level was generally low but at almost regular intervals would increase at least fivefold, persist for a few cycles, and then resume its low level. These bursts would repeat themselves. At low voltages a normal noise pattern similar to that of Fig. 6 (c and d) appears.

In the early work, before these various effects were completely understood, many seeming anomalies were noted, such as apparent failure of the noise to occur with certain types of metal electrodes. We now believe that the single most important variable is the relative humidity at which the sample has been stored. In general the intensity of noise increases with the moisture content and polarity of the sample.

Figure 3 illustrates the effect of relative humidity on coated Cellophane. At zero percent R.H. the sample showed very little activity even at 300 volts and the few displacements which were observed were mostly in one direction with respect to the base line of the recording. At 50 percent relative humidity and with the voltage decreased to 150 volts, the noise has increased tremendously, and the pattern is distributed on both sides of the zero. Figure 4 shows similar behavior for a copolymer of 60 percent vinylidene chloride—40 percent acrylonitrile where the humidity has been carried to 86 percent. At 86 percent R.H. and 250 volts, the loudspeaker was emitting a siren-like noise of gradually increasing pitch and intensity. Figure 5 shows the effect of relative humidity on Nylon; in this case, at 86 percent R.H., the voltage was dropped to 67 volts and still gave a pronounced effect. Here again it is particularly evident that the character of the noise changes markedly on increasing the voltage: low voltages give records symmetrical about the zero line, high voltages gave a more assymmetric recording. This effect will be discussed in the theoretical section.

It should be emphasized that the relative humidity values just given refer to those at which the samples had been stored for 24 hours or more. The measuring electrodes (with associated resistances and capacitors) were housed in a brass case for good electrical shielding. When

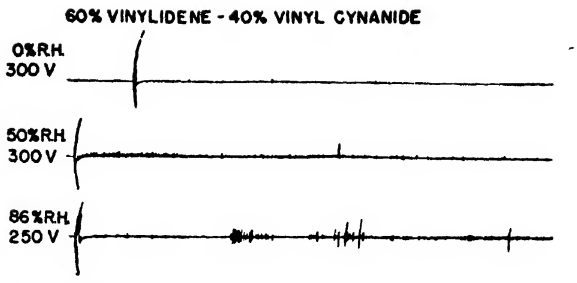


FIG. 4. Effect of relative humidity on current fluctuations in a 60 percent vinylidenechloride—40 percent vinylcyanide copolymer. The fluctuations are unsymmetrical about the axis at high voltage, become symmetrical at higher humidities and lower voltages. The loud speaker emitted a siren-like noise while the bottom line was being recorded.

the film sample was removed from its humidity chamber and placed in the measuring apparatus, where ambient conditions prevailed, the sample either lost or gained moisture. This fact accounts for some but not all of the anomalous behavior which was found. Since the relative humidity of the room was around 20–30 percent for many of these measurements, the very active samples of high moisture content tended to dry out and lose activity. However, Cellophane has an equilibrium moisture content of about five percent even at 20 percent R.H. and hence still has fairly high activity at all times. We noted one case in which the voltage remained on a Cellophane sample overnight; in the morning the sample showed no noise until the voltage was reversed, when it was extremely noisy.

We might conclude this section with a brief description of several additional experiments. Various types of electrodes were tried, such as aluminum, brass, lead and silver, with no appreciable differences as long as a fresh sample of plastic was taken each time from the humidity chamber. Brass was the standard electrode material and was used for all curves given in this report. In general a one-inch diameter electrode was used unless otherwise noted.

A sevenfold range of sample thickness (0.0004 to 0.0028 in.) at a constant potential gradient of 15 volts per mil on a 60 percent vinylidene chloride—40 percent acrylonitrile copolymer at 50 percent R.H. was studied. Each sample was run on one polarity until the noise

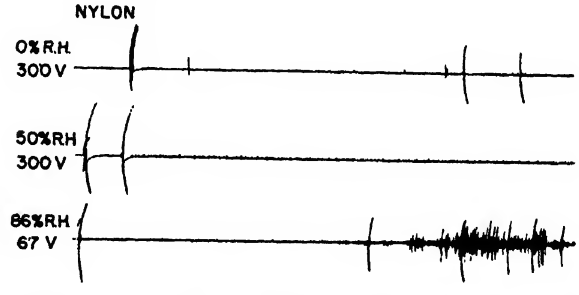


FIG. 5. Effect of relative humidity on Nylon noise pattern. The time scale here as in Figs. 3 and 4 is one minute per horizontal traverse.

stopped. This end point was naturally somewhat arbitrary, but in general the duration of the noise increased linearly with the thickness.

Some of the more noisy materials such as Cellophane were run between two films of polystyrene. They still gave a characteristic noise diminished in intensity. Polystyrene alone did not show this effect unless it had been conditioned first at 86 percent R.H. Films of water soluble methylcellulose were run but did not give any noise apparently because of their extremely high conductivity. If, however, a methocel film was sandwiched between two sheets of polystyrene, it did give a reproducible pattern both on applying and removing the potential, although this pattern does not have the character of a noise but rather of an oscillatory discharge.

We are reserving until a later time studies of the effect of temperature on these noise phenomena. It will be necessary to vary the temperature while keeping the moisture content of the sample approximately constant.

Before concluding this experimental section we wish to call attention to some of the prior literature where analogous effects have been described. Haworth and Bozorth[3] in studying pre-breakdown phenomena in glass observed that the current is "noisy," and occurs in pulses which grow in magnitude as the field intensity increases. Von Hippel[4] has also observed similar effects. Tykociner, Brown, and Paine[5] observed oscillations associated with oil-impregnated paper at high field strengths, and arising from voids in the dielectric. Austen and Hackett[6] ascribe similar effects to the sudden discharge of a local region of the dielectric of lower electric strength than the dielectric in series with it: This discharge may give the appearance of an oscillating current if the circuit constants are of the correct magnitude. Quite recently Glass[7] has studied leakage noise in a number of dielectric materials, both ceramic and organic, at high humidities and high applied voltages. His principle source of noise was surface leakage at the high relative humidity of the samples. Teflon and fluorothene were the only materials free from such effects.

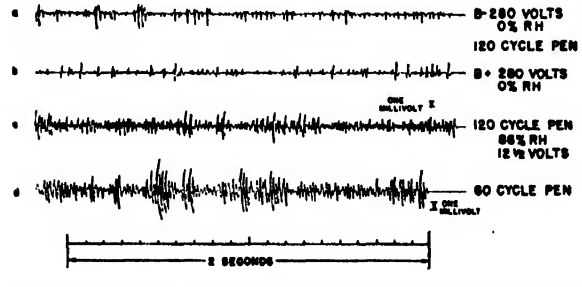


FIG. 6. Fast chart speed record showing detailed noise patterns for Nylon. Lines a and b show unsymmetrical surges at high voltage and low humidity, using the 120-cycle pen. Lines c and d show the symmetrical pattern at low voltage and high moisture content.

Seitz has given a general theory[8] of electron multiplication processes in crystals, which lead to the current avalanches that precede and eventually cause dielectric breakdown. The effects described in the present paper occur at voltages from 15 to 300 volts per mil, whereas the dielectric strengths of the various materials studied range from about 500 to 5000 volts per mil.

Incidentally, some years ago Whitehead and Nethercot[9] studied dielectric breakdown in cellulose acetate, which they called a moist dielectric. They found that this material did not obey Ohm's law but that the conductivity increased linearly with voltage up to 140,000 volts/cm, and even more rapidly at still higher voltages. They spoke of "ion avalanches" as occurring at these higher voltages and actually preceding dielectric breakdown.

More recently Whitehead[10] has presented oscillograph tracings showing the current resulting from applying an alternating potential to polyethylene. There is good evidence on these tracings for what we would call a noise pattern.

THEORETICAL DISCUSSION

The general effects just described in Figs. 2–5 might arise from surface or volume ionic conduction or from rotation of large regions of dipoles. It seems quite possible that several of these effects might be operating together. We are inclined to rule out dipole rotation first on the basis of some of the long-time effects observed, and secondly, because of the marked effect of moisture on polymers such as saran which one would not expect to be markedly plasticized by the absorption of a few hundredths of a percent of water. The cooperative movement of large groups of dipoles would likely be confined to crystalline polymers. Surface conduction, while possibly important at high relative humidities, cannot be of any appreciable importance for samples stored at zero percent R.H., where the noise is still observable at high voltages. We did not use a guard ring on the electrodes. Surface conduction would certainly follow a different mechanism than that of volume conduction, and probably would not place as great dependence on the size of the diffusing ion.

We are thus strongly inclined to view the various phenomena discussed in this paper as arising from the migration of ions through the polymer under the applied electric field. The ions are presumably those of H^+ and OH^- although in the case of saran it is quite possible that Cl^- ions are also present while each of the other polymers may have characteristic impurity ions, both organic and inorganic.

[3] F. E. Haworth and R. M. Bozorth, Physics 5, 15 (1934).
[4] A. R. von Hippel, Phys. Rev. 54, 1096 (1938).
[5] Tykociner, Brown, and Paine, University of Illinois, Eng. Exp. Station Bulletin, Nos. 259 and 260 (1933).
[6] A. E. Austen and W. Hackett, J. Inst. Elec. Eng. 91, I 298 (1944).
[7] F. M. Glass, Rev. Sci. Inst. 20, 239 (1949).
[8] F. Seitz, Phys. Rev. 76, 1376 (1949).
[9] S. Whitehead and W. Nethercot, Proc. Phys. Soc. 47, 974 (1935).
[10] S. Whitehead, Trans. Faraday Soc. 42A, 66 (1946).

According to Stearn and Eyring,[11,12] the migration of ions in an electric field is susceptible to the usual absolute reaction rate treatment whereby the observed ionic currents are the net resultant of a forward movement of the ions in the direction of the applied field and a backward diffusion because of the concentration gradient of ions thus established. Or

$$\text{Net rate} = (NDc\varphi ze/kT) - NDdc/dx. \qquad (1)$$

N = Avogadro number. D = Diffusion constant of ion. c = Concentration of ion. φ = Potential gradient. z = Valence of ion. e = Electronic charge. dc/dx = Concentration gradient. k = Boltzmann constant. T = Absolute temperature.

Under high potential gradients the second term is almost negligible and the observed current fluctuations are primarily in one direction as may be seen in Fig. 6(a and b) for Nylon film of 0.9 mil thickness at zero percent R.H. and 280 volts applied potential. At higher relative humidities where the fluctuations are much greater and the voltages are greatly reduced, then these fluctuations occur on either side of zero, as can be seen in Fig. 6(c and d) for the same Nylon film.

It might be pointed out that these current fluctuations are frequently of such magnitude as to cause an observable shift in the "steady" current. This can be seen, for example, in line 3 of Fig. 2 where the entire noise pattern shifts its center. Such effects can also be seen on the cathode-ray oscillograph where the base line moves up or down with large fluctuations. It would be desirable, obviously, to record this d.c. component simultaneously with the noise pattern.

We believe these fluctuations arise because of the nature of the medium in which the ionic conduction occurs. One is quite familiar with the mechanism for moisture vapor transmission through polymers as pictured by Barrer and others in terms of absolute reaction rate theory. This theory pictures that as a result of segmental motion of the polymer chains holes occasionally open through which the water molecules can diffuse. This segmental motion presumably has a natural frequency corresponding to that of the peak in the loss factor *vs.* frequency curve for various polymers.

If we apply a similar picture to ionic migration, it will be recognized that on a molecular scale the movement of ions in the direction of the field will be sporadic and that fluctuation phenomena might occur at about the usual frequency for dielectric response. Thus the highly viscous nature of the polymeric medium presumably slows down hole formation to a convenient rate for observation, and makes the measurement of conduction current through such a specimen a convenient tool for the study of this internal molecular process.

If, however, the energy acquired by an ion in the applied field is not too large as compared with kT, then the backward jumps of the ions, as indicated by Eq. (1) will become important. This backward rate must also occur at a rate governed by the frequency of motion of the chain segments. Moreover, it seems likely, especially if H^+ and OH^- ions only are involved, that both types of ions will be subject to the forward and backward jumps, each at its own characteristic rate. Thus the observed noise will be a composite resultant of the random motions

TABLE I. Distribution of amplitudes of noise.

Ordinate in mm	No. of intersections on 60-cycle pen tracing: Cellophane 180 v 50% R.H. 5 mm = 10 mv	Nylon 200 v 50% R.H. 5 mm = 1 mv	Nylon 12.5 v 86% R.H. 5 mm = 1 mv
0	2	2	2
1	4	6	4
2	6	10	8
3	6	12	12
4	8	14	14
5	14	21	16
6	18	25	25
7	26	32	33
8	38	39	38
9	62	44	62
10	88	52	74
11	94	58	88
12	114	70	101
13	102	70	111
14	86	63	102
15	70	63	104
16	64	58	88
17	40	50	68
18	24	45	56
19	18	34	48
20	16	30	34
21	6	20	24
22	2	14	14
23	2	14	10
24	2	6	7
25	—	4	4
26	—	2	2
27	—	2	—
	942	860	1149
Time of observation	1 sec.	0.6 sec.	1 sec.
Root-mean-square deflection	3.9 mm	5.4 mm	4.5 mm
Root-mean-square voltage deviation Δv_{rms}	7.8 mv	1.08 mv	0.9 mv
Root-mean-square current deviation Δi_{rms}, in electronic charge units	12.5×10^9	1.8×10^9	1.4×10^9
Resistance of sample, ohms	1.5×10^8	1.5×10^9	5×10^7
Calculated average current, i_0, in electronic charge units	8×10^{12}	10^{12}	1.6×10^{12}
$(ei_0/t)^{\frac{1}{2}}$	2.8×10^7	10^7	1.3×10^7
Ratio: $\Delta i_{rms}/(ei_0/t)^{\frac{1}{2}}$	4.5×10^2	1.8×10^2	1.1×10^2
Ratio: $\Delta i_{rms}/i_0$	1.5×10^{-3}	1.8×10^{-3}	1×10^{-3}

[11] A. E. Stearn and H. Eyring, J. Phys. Chem. 44, 976 (1940).
[12] Glasstone, Laidler, and Eyring, The Theory of Rate Processes, (McGraw-Hill Book Company, Inc., New York, 1941), p. 552 ff.

of the ions, which motion depends on the internal mobility of the system, on the size of the ions, on the temperature, and on the applied field.

Some numerical calculations to illustrate orders of magnitude may be helpful at this point. For example, a potential of 250 volts applied to a 1-mil film is 100,000 v/cm which corresponds to an energy of 1.6×10^{-7} erg per cm. Assuming that an ion jumps a distance of 5×10^{-8} cm each time, it acquires an energy of 8×10^{-15} erg. kT is 4×10^{-14} erg and hence at 250 volts the electrical energy is 20 percent of kT. Thus the energies we are using vary from about 1 to 20 percent of kT which accounts already for the characteristic difference in the patterns of Fig. 6(a and c).

Next some estimate of the magnitude of these current fluctuations may be of interest. Thus, in a typical example, Nylon film at 50 percent R.H. and 200 volts gave an extreme fluctuation range of 5 mv. From the nature of the circuit one can calculate that this would correspond to about 10^{-10} electronic charges per burst. Many of the smaller bursts would involve only a tenth of this amount. The steady current, calculated from specific resistivity, is about 10^{12} electronic charges. Additional calculations are given in Table I.

We wish to return now to Fig. 6, which represents studies on nylon at high and low voltages at full chart speed. Lines a and b represent zero percent R.H. and 280 volts with reversal of polarity. In line a the strong surges are seen to be going down with only minor surges in the upward direction; whereas in line b with reversed polarity the strong surges are upward. This conductivity in Nylon at zero percent relative humidity may arise from mobile hydrogen atoms, according to the mechanism suggested by Baker and Yager.[12a] Lines c and d at 86 percent R.H. and 12½ volts illustrate the surges uniformly distributed on either side of the zero line. Line c was recorded with the 120-cycle pen which shows more fine detail than does line d which is on the 60-cycle pen. These two traces are shifted slightly with respect to each other (as can be seen at the right hand side) because of different lengths of the pen arms. It is possible to note the same events on each record and to see that while they do have the same general features yet they actually differ in details.

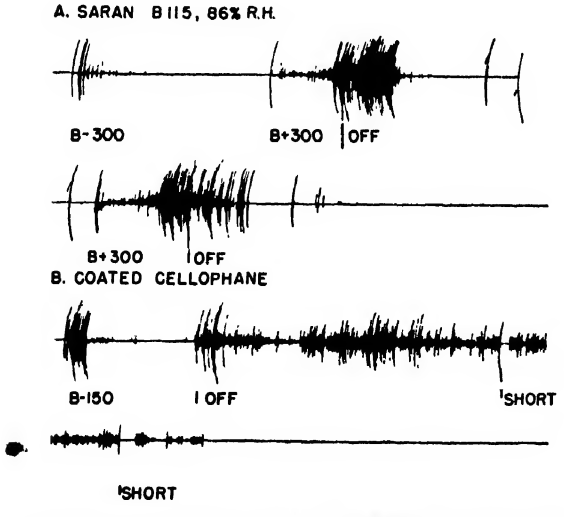


FIG. 7. Typical noise patterns which persist after the voltage is removed from the samples. A is regular saran film which had 300 volts ($B-$) applied, then polarity reverse ($B+$) and voltage off. The same sequence is repeated in the second line. Similar results are shown for coated Cellophane at 86 percent R.H., except that the noise lasts much longer. The two points marked "short" signify that the amplifier input was shorted.

The curves of Fig. 6(c and d) illustrate a principle which is common in fluctuation phenomena, namely, that when the disturbing force is small (low voltage in this case) the fluctuations appear to follow roughly the natural frequency of the system. Thus 6c has about a frequency of 120 cycles, while line 6d corresponds more closely to 60 cycles. Kappler observed this type of behavior with galvanometer mirrors when the ambient pressure was reduced to a low value.[12b] This behavior does not affect the root-mean-square deviation.

Tracings of the kind shown in c and d were obtained both on Nylon and coated Cellophane, using a cross-sectioned recording paper having one millimeter squares supplied with the recorder. These records facilitated statistical studies of the nature of the noise record. For example, by counting the number of intersections of lines parallel to the time axis with the curve, as recommended by Furth[13] and recently applied by Furth and MacDonald,[14] we obtained the distribution data listed in Table I. The 60-cycle pen tracing was used in each case, and the bottom-most point reached by any tracing was taken as zero millimeters for the ordinate scale. It was a coincidence that each of these occupied roughly 25 mm since the actual voltage range was 10-fold greater for the Cellophane. Finally, a similar statistical study was made of the events in line d of Fig. 6, for a period of one second.

It is evident from the distribution data of Table I that the noise peaks are symmetrical about the zero line even up to 200 volts applied potential under conditions when the average current through the film tends to be large. Then the second term of Eq. (1), Ddc/dx, is apparently large and the backward current is appreciable. If, however, as indicated by lines a and b of Fig. 6 the voltage is high and the moisture content low, ideal conditions prevail for observing unsymmetrical surges. In other words, there are presumably so few ions present that the concentration gradient, dc/dx, of ions cannot acquire a large value.

We must now examine in more detail the question of the random nature of these fluctuation currents.[14a] We

[12a] W. O. Baker and W. A. Yager, J. Am. Chem. Soc. 64, 2177 (1942).
[12b] E. Kappler, Ann. I. Physik 11, 233 (1931).
[13] R. Furth, Ann. d. Physik 53, 177 (1917).
[14] R. Furth and D. K. C. MacDonald, Proc. Phys. Soc. 59, 388 (1947).
[14a] R. B. Barnes and S. Silverman, Rev. Mod. Phys. 6, 162 (1934).

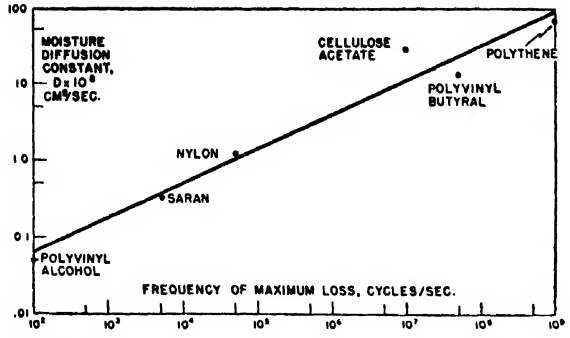


FIG. 8. Diffusion constant for moisture vapor as a function of the frequency at which the electrical loss factor of the polymer is a maximum.

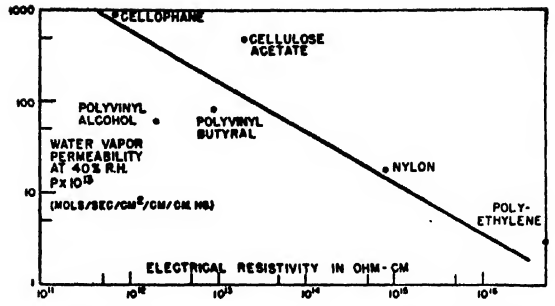


FIG. 9. Correlation between moisture vapor permeability (which is the product of diffusion constant and solubility) and the direct current resistivity of a series of polymers.

have been assuming that the conduction current is composed only of current pulses arising from ion jumps in the dielectric medium, and that these jumps occur at random. If we observe the net current (or a potential drop across a resistance) in the time interval, t, then the root-mean-square current fluctuation, Δi_{rms} is

$$\Delta i_{rms} = (ei_0/t)^{\frac{1}{2}}. \quad (2)$$

Where e is the electronic charge in coulombs, i_0 is the average current, and t is the time interval. t is, for the 60-cycle Brush oscillograph pen, approximately 0.01 sec. Hence, for a typical case i_0 is 10^{-9} amp. or 10^{10} electronic charges per second. For this case one calculates $\Delta i_{rms} = 10^5$ electronic charge units per second. Alternately, one may use the Schottky expression

$$\Delta i_{rms} = (2ei_0\Delta F)^{\frac{1}{2}} \quad (3)$$

where ΔF is the band width and the other symbols have their previous significance. Assuming ΔF as 100 cycles, then the numerical value used above of $i_0 = 10^{-9}$ amp. gives $\Delta i_{rms} = 1.4 \times 10^6$ electronic charges per second, which differs from the preceding value only by the factor $(2)^{\frac{1}{2}}$. The several average and statistical quantities are given in the appropriate columns of Table I. It is seen that Δi_{rms} is from 110 to 450 times the value one would calculate from the average current if normal shot effect conditions applied. Also Δi_{rms} is seen to vary from 0.2 to 5 percent of the steady current. It must be re-emphasized that these values are probably indicative only until we can make simultaneous measurements of the actual steady current.

The fact that these Δi_{rms} values are at least a hundred times larger than one would calculate from the steady current requires a revision of our original premises. The simplest revision is to assume that groups of ions move randomly in the form of "avalanches." Normally the term "avalanche" implies the cumulative building up of a large number of ions by successive collisions. Such a meaning is not intended here. Rather, it seems more likely that if a "hole" opens through the dielectric, it may persist long enough to permit a large number of ions to move through. This may be particularly marked at high moisture contents where small pockets of water exist within the dielectric. It is further possible that the passage of a few ions between a certain pair of chain segments may generate sufficient local heat to keep a path open for additional ions.

Moreover, there is an additional aspect of Eq. (1) which has unique significance for a polymeric dielectric. In any normal electrolysis experiment it does not seem likely that appreciable concentration gradients of ions will exist. However, trapping mechanisms in a high polymer may well give rise to some extremely large concentration gradients. Also, the average current, i_0, which one observes consists of the difference between a forward and backward current,

$$i_0 = i_f - i_b. \quad (4)$$

It is conceivable that the two directional currents, i_f and i_b, could both be large with respect to i_0. The fluctuating current should be based on these individual currents i_f and i_b. Hence the observed fluctuations in current may well seem extremely large in comparison to the average current.

AFTER EFFECTS

There is one final effect which should be mentioned: This noise will frequently persist for a few seconds after the voltage has been removed from the sample. Figure 7 shows some typical records.

According to Maxwell's rule, the relaxation time, τ, for an arbitrary distribution of charge in a solid, is given by $\tau = \epsilon'/\sigma$ where ϵ' is the dielectric constant and σ is the conductivity in m k sec units. Thus a material of dielectric constant 5 and resistivity of 10^{12} ohm-cm should have a relaxation time of 0.4 sec. Some of these observed times are of the order of one minute. Pollissar has discussed conduction in liquids on the basis of a cage model.[15] It seems that in a solid plastic, particularly one that can crystallize, the concept of ions trapped in local cages acquires a high probability. It may well account for some of the long-time constants observed here after the voltage is removed. A highly viscous medium, and especially one in which motion of the ions proceeds in discrete steps, is an ideal one for trapping ions. Two

[15] M. J. Polissar, J. Chem. Phys. 6, 833 (1938).

TABLE II. Correlation of moisture vapor transmission with dielectric properties of plastics.

Material	Moisture[a] diffusion constant $D\times10^9$ in cm²/sec.	Moisture[a] vapor permeability $\times10^{13}$ in moles/sec. cm²/cm/unit pressure differential	Frequency[b] of power factor maximum in cycles/sec.	D.C.[b,c] resistivity in ohm-cm
Polyvinyl-alcohol	0.05	60	$<10^3$	2×10^{12}
Vinylidene-chloride-acrylonitrile copolymer	0.32	0.7	10^3	7×10^{16}
Nylon	1.2	18	5×10^4	8×10^{14}
Polyvinyl-butyral	13	83	4×10^7	9×10^{12}
Cellulose acetate	30	500	10^7	2×10^{13}
Polyethylene	68	2.9	10^9	$>5\times10^{16}$
Cellophane		900		7×10^{11}

[a] From Hauser and McLaren at 40 percent relative humidity.
[b] From von Hippel.
[c] At 30 percent R.H.

pockets of ions of opposite sign may either move further in the paths toward their respective electrodes, or may mutually discharge each other, giving rise to considerable current surges in the backward direction. It might be noted that the sudden stoppage of an "avalanche" of ions would have the appearance of a surge in the reverse direction.

An inspection of the circuit will indicate that the effective time constant becomes much longer when the potential is removed, based on the discharge of the 0.001 mica condenser to ground through the sample. Thus the time constant can be of the order of several seconds, depending on the actual leakage resistance of the experimental condenser. However, some of these persistent aftereffects last much longer and are the result, we believe, of trapped ions.

RELATIONSHIP OF THIS NOISE TO MOISTURE VAPOR TRANSMISSION

If our picture of this noise is correct: i.e., arising from the activated diffusion of ions where the rate-controlling step is the speed with which chain segments can rotate, then there should be an intimate correlation between these noise effects and moisture vapor transmission. According to Barrer[16–19] the transmission of water through a plastic membrane is given by the expression

$$P = D \cdot S \qquad (5)$$

where D is the diffusion constant in cm²/sec. and S is the solubility of the water molecule in the polymer medium while P is the observed permeability constant in moles per cm² per sec. per unit thickness per unit differential water vapor pressure. Now the diffusion rate should be controlled, among other factors, by the speed with which the segments can open. Hence one would expect a direct relationship between the diffusion constant, D, and the frequency of the maximum in the dielectric response curve, i.e., the frequency at which the loss factor is a maximum. Figure 8 based on the moisture diffusion data of Hauser and McLaren[20] and the dielectric data of von Hippel and associates[21] shows that this is approximately true. Polyethylene itself has no dipoles to give rise to loss so that the observed value of 10^9 cycles presumably arises from impurity groups such as carbonyl. It will be recognized of course that each dielectric material has a differently shaped loss factor *vs.* frequency curve and that the length of the segment engaging in the dielectric response will vary. In some cases also it is mainly a side group rather than a segment of the main valence chain which responds to the applied electric field. Moreover, in Fig. 8 the slope is 0.5, indicating that D increases only as the square root of the segmental frequency.

Nevertheless, we believe the general features of Fig. 8 to be essentially correct and extremely illuminating. It is seen from Table II that Saran and polyethylene represent extreme cases as regards diffusion even though they are relatively close on the final moisture vapor transmission. The outstandingly low moisture vapor transmission of Saran would, in this point of view, be ascribed to the fact that the large chlorine groups have slowed down chain segment rotation to a relatively low frequency. Polyvinyl alcohol is an even more extreme case than saran as far as diffusion at low relative humidities, and here it is presumably both the size of the OH group and hydrogen bonding which slows down dielectric response.

We might mention that results similar to those of Fig. 8 also hold for the diffusion of nitrogen and butane through a series of vulcanized rubbers, using the diffusion data of Barrer and Skirrow[19] and the dielectric data of Scott, McPherson, and Curtis.[22] So far as we know, the seemingly obvious relation between diffusion and dielectric response frequency has not hitherto been noted. However, Levi[23] has already noted that the activation energies and entropies for dielectric relaxation and for moisture vapor transmission follow essentially the same pattern. A further extension of Fig. 8 would be to seek a correlation between the noise frequency spectrum and the dielectric response curve. We have observed qualitatively, on a fast cathode-ray oscillograph, that polyethylene does have higher frequence noise effects than does saran.

Next, if we return to a consideration of Eq. (1), we see from the first term that the electrical conductivity arising from a direct current field is proportional to the product $D \cdot c$ where D is the diffusion constant of the ion

[16] R. M. Barrer, Trans. Faraday Soc. 34, 849 (1938); 35, 628, 644 (1939); 36, 1235 (1940).
[17] R. M. Barrer, Diffusion in and through Solids (Cambridge University Press, New York City, 1941).
[18] R. M. Barrer and G. Skirrow, J. Polymer Sci. 3, 549 (1948).
[19] P. Doty, J. Chem. Phys. 14, 244 (1946).
[20] P. M. Hauser and A. D. McLaren, Ind. Eng. Chem. 40, 112 (1948).
[21] A. R. von Hippel, "Tables of dielectric materials," Laboratory for Insulation Research, M.I.T., Vol. II (June, 1945).
[22] Scott, McPherson, and Curtis, Bur. Stand. J. Research 11, 173 (1933).
[23] D. L. Levi, Trans. Faraday Soc. 42A, 152 (1946).

and c is the concentration of ions. But these are precisely the two factors on which moisture vapor permeability depend: except that moisture diffusion refers to the whole electrolyte molecule rather than its constituent ions. However, there is an intimate relation between the diffusion of an electrolyte and the diffusion of its constituent ions. Hence, one is led to predict a relationship between the moisture vapor permeability of polymer films and their direct current conductivities or resistivities. Figure 9, using the data collected in Table II, shows that this relationship holds surprisingly well.

Actually, in the usual electrical resistivity test the sample has been equilibrated in an atmosphere of constant humidity and has a uniform water content throughout its thickness; in a moisture vapor test the water content varies through the film. In addition, different samples were used for the two types of tests. Also, various impurity ions may be present from sample to sample. These various uncertainties may account for the fact that the slope of Fig. 9 is not quite unity.

One may ask now whether or not water vapor diffusion proceeds in part or entirely by an ionic mechanism. For example, a hydrogen ion, having a higher mobility, might diffuse forward and then pull along the slower $-OH$ ion. This is probably the source of the electric currents which Fischer and Muller[24] found to accompany the diffusion of water through plastic films until the steady state was reached. They found that the electrical current passed through a maximum whose height was greater the larger the moisture vapor permeability of the film, and then fell to zero under steady-state moisture transmission. Since the net charge transport would be zero such currents would depend on the difference in velocities of the hydrogen and hydroxyl ions.

[24] C. Fischer and F. Horst Muller, Kolloid Zeits. **101**, 43 (1942).

SUMMARY

This report is concerned with a description and preliminary interpretation of random fluctuation currents which accompany the application to thin films of moist dielectrics voltage gradients which are higher than those used in normal electrolysis experiments and lower than those leading to dielectric breakdown. These fluctuating currents are believed to arise from the fact that both the forward motion of the ions in the direction of the field and their backward movement because of concentration gradients thus established are regulated by the segmental motion of the polymer chains. It appears now, at least for the low frequency part of the noise spectrum which we analyzed, that groups of ions rather than single ions are acting as the random charge carriers.

This present report must be recognized as a preliminary study of the problem. The following are some of the many experiments which must be performed to elucidate fine details: effect of temperature; simultaneous recording of the average current, especially as it depends on applied voltage; study of the noise frequency spectrum; evaluation of surface *vs.* volume conductivity effects; a study of various ion types other than those of water, such as ammonium, chloride, etc.; and a more detailed study of aftereffects. Finally, a more complete theory needs to be developed. It is interesting to note that Richardson has developed a theory of fluctuations arising from diffusional mechanisms to account for noise developed at electrical contacts.[25]

It is a pleasure to acknowledge the assistance of Herbert Bauss who built the apparatus and aided in making the observations; to F. Towsley for discussions about noise problems in electronic circuits; and to E. B. Baker and L. A. Matheson for helpful discussions about the general significance of these results.

[25] J. M. Richardson, Bell Sys. Tech. J. **29**, 117 (1950).

Emulsion Polymerization with Ultrasonic Vibration

A. S. Ostroski and R. B. Stambaugh
*The Goodyear Tire and Rubber Company, Akron, Ohio**

Emulsion-type polymerization reactions have been accelerated by irradiation with ultrasonic energy.

A magnetostriction oscillator at 15 kc gave essentially the same results as a piezoelectric oscillator at 500 kc. The rate of formation of polystyrene was doubled and higher final yields obtained using about 0.03 watt of vibrational power per cc of latex. Experiments were made to separate local heating effects from those due to vibrational energy.

A thermostatted stainless steel cell was developed for irradiating butadiene emulsions under pressures up to 60 p.s.i. About ½ watt of vibrational power per cc of latex doubled the rate of reaction for GR-S, Redox, and Redsol formulas.

The ultrasonic intensity apparently must exceed a critical value before any appreciable accelerative effect is observed. Above the critical intensity, the time to reach a given yield of polymer is approximately inversely proportional to the power used.

I. INTRODUCTION

WHILE the more spectacular phenomena associated with intense ultrasonic radiation have been described by many observers, very little information about practical applications of this energy has been published.[1] Colloidal systems[2] may be affected as to rate of emulsification, particle size and stability, and certain chemical reactions[2,3] are accelerated by the action of ultrasonic waves. The breakdown or de-polymerization of polymers in solution has been studied,[1,4] and the polymerization of acetic acid by treatment with ultrasonic power has been reported.[1] Apparently nothing has been published on the effects of ultrasonic vibration on emulsion polymerization systems such as are used in the production of synthetic rubber and with which the present investigation is concerned.

II. MAGNETOSTRICTION APPARATUS

Both a magnetostriction type of oscillator, operating at 15 kc, and a piezoelectric oscillator, operating at 500 kc, were used as sources of vibrational energy. The magnetostriction oscillator consisted of a laminated nickel bar, 6 in. long and ¾ in. square, excited by a coil made up of 28 turns of ⅛-in. copper tubing through which water was forced for cooling. The coil carried both a d.c. polarizing current and an a.c. current of about 15 to 20 amp. at 400 volts, the latter being supplied by a variable frequency vacuum tube oscillator capable of about 1000 watts output. The bar was supported by a spider inside the insulated coil, and the assembly immersed in the emulsion to be treated. The liquid was maintained at the desired temperature by rapid stirring and the circulation of thermostatically controlled fluid in the jacketed walls of the stainless steel vessel. The chamber was covered and a stream of nitrogen was piped in to prevent contamination by oxygen. A block diagram of this system is shown in Fig. 1, and a photograph of the apparatus is shown in Fig. 2.

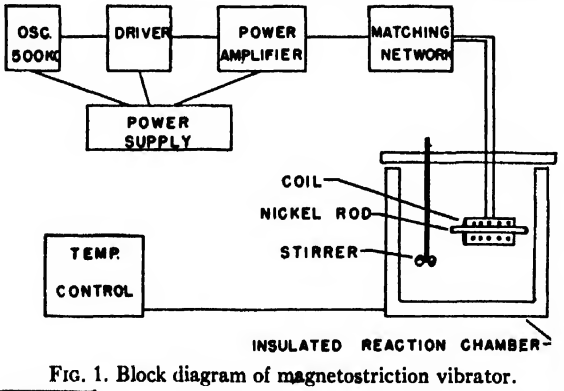


FIG. 1. Block diagram of magnetostriction vibrator.

The reaction was run at atmospheric pressure, and a simple styrene formula was used, containing: 100 parts (by wt.) of styrene monomer; 180 parts, distilled water; 5 parts, soap; 0.6 part, potassium persulfate. Samples were extracted by means of a pipette at intervals during each run so that the course of the reaction could be followed. The polystyrene in each sample was precipitated out in alcohol, dried, and weighed. Figure 3 shows some typical curves obtained with the bath at 50°C with 600- and 1200-cc batches. The percent of the original monomer transformed into polymer (yield) is plotted against the time elapsing after the styrene and soap solution, separately preheated to the bath tempera-

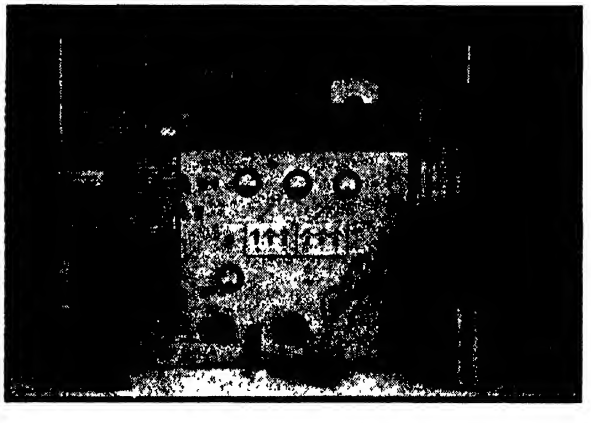


FIG. 2. Photograph of magnetostriction system.

*Contribution No. 174 from the Goodyear Tire and Rubber Company Research Laboratory.

[1] A list of references is given by Karl Sollner in his chapter on "Sonic and ultrasonic waves in colloid chemistry" in *Colloid Chemistry* (Reinhold Publishing Corporation, New York, 1944), edited by J. Alexander, Vol. 5. Also see D. Thompson, Chem. Eng. Progress 46, 3 (1950).

[2] R. W. Wood and A. L. Loomis, Phil. Mag. 4 (7), 417 (1927).

[3] A. Kling and R. Kling, Comptes Rendus 223, 33 (1946).

[4] H. Mark, J. Acous. Soc. Am. 16, 3, 183 (1945).

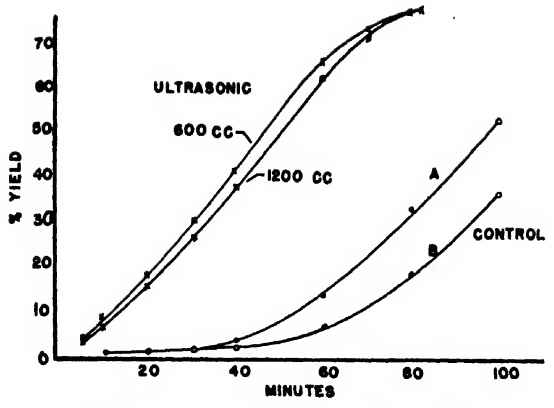


FIG. 3. Conversion of styrene at 50°C, 15 kc.

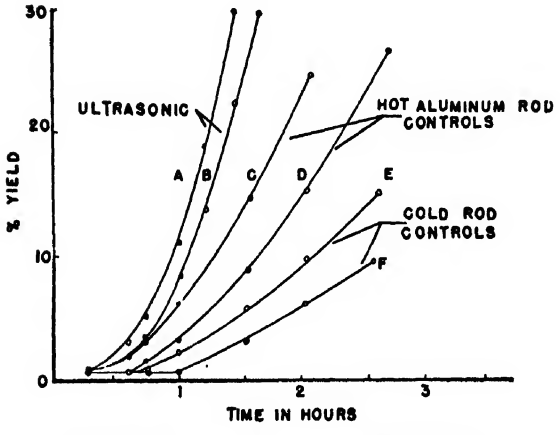


FIG. 4. Conversion of styrene at 40°C, 15 kc.

ture, were poured together into the vessel. Controls were run in the same apparatus without current in the coil.

The principle effect of the ultrasonic agitation appears to be the elimination of the induction period, with only a slight increase in the actual rate of the reaction as given by the slope of the curves.

At 40°C, the effect of the irradiation is much more pronounced. As illustrated in Fig. 4, the two curves A and B, obtained with the nickel rod excited, show a much more rapid conversion of the styrene than for curves E and F, when the rod was not energized. Due to magnetic and elastic hysteresis and eddy current losses, a great deal of heat is generated in the rod during a run. This results in a higher temperature in the emulsion immediately around the rod even though rapid stirring is maintained.

In order to show that the acceleration of the reaction was not due solely to this temperature rise, an aluminum rod was substituted for the nickel rod. This was heated by eddy currents to the same temperature as reached by the magnetostriction rod during a run, as measured by thermocouples welded to the surface of the middle section. Runs made with this hot aluminum rod produced curves C and D which show that the rod temperature does have some effect, but not sufficient to explain the rapid reaction of curves A and B which must be attributed to the ultrasonic agitation.

The nickel bar is very inefficient as a source of vibrational energy. The best estimates of actual acoustic power obtained with a kilowatt input to the amplifier were about 50 watts. This fact, combined with the difficulty of designing an enclosed system which could be operated under a relatively high pressure, made it desirable to use a piezoelectric crystal for further measurements.

III. QUARTZ OSCILLATOR

An X-cut quartz crystal was used for a source of energy at 500 kc, powered by a radiofrequency oscillator having a plate input of up to 300 watts. The crystal was 3.81 cm in diameter, having about 8 sq. cm of usable radiating area in contact with the emulsion. An estimated 100 watts of ultrasonic energy was realized with 200 watts plate input to the oscillator which was the usual power level. A few runs were made with 300 watts plate input, but several crystals were shattered in attempts to go above this level.

The crystal was mounted in the special reaction cell pictured in Fig. 5 which could be run at internal pressures of 60 lb./sq. in. This consisted of a stainless steel block through which a 2-in. diameter hole was drilled for the reaction chamber. A number of $\frac{1}{4}$-in. holes were drilled parallel to this chamber for circulation of temperature controlling fluid. The crystal was mounted against a hole cut through the bottom of the reaction

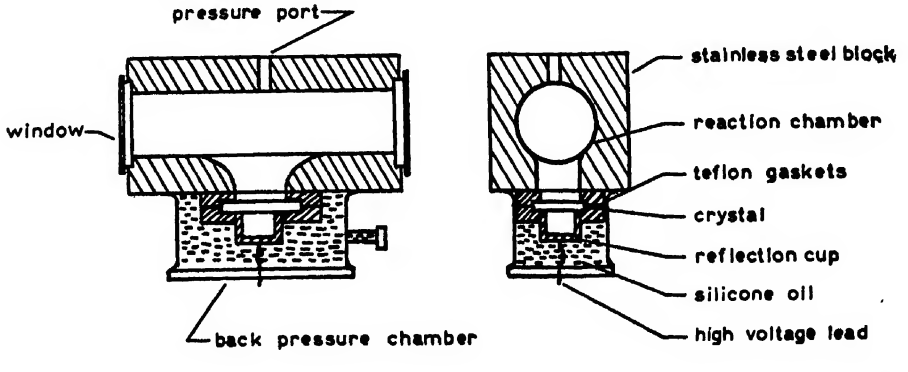


FIG. 5. Stainless steel reaction cell.

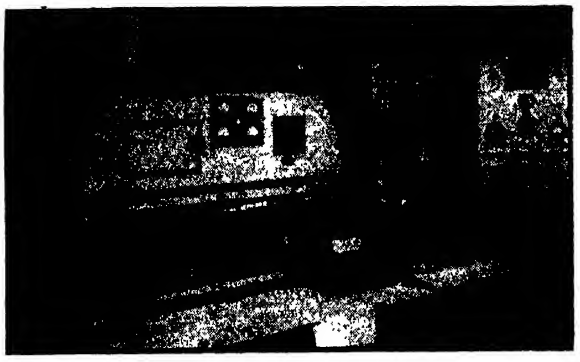


FIG. 6. Photograph of piezoelectric system.

chamber. It was held in place and insulated with Teflon gaskets. The high voltage side of the crystal was surrounded with silicone oil in a steel chamber welded to the block, having a polystyrene lid bolted to a flange on the bottom. The high voltage lead passed through this lid and made a spring contact to the brass, air-filled back-reflection cup which contacted the silvered back face of the crystal.

In the top of the block, a stainless steel pipe and valve was screwed into a threaded hole for filling and emptying the chamber and for insertion of a thermocouple. A pressure gauge was mounted on a side arm off this pipe. Heavy glass windows at each end of the cell provided means for observing the emulsion while it was being irradiated. These could be removed for cleaning the cell.

The photograph of Fig. 6 shows this apparatus with the associated recording and controlling systems for temperature, pressure, and power. The reaction cell in the center foreground is mounted in a cradle with a pneumatic piston beside it which rocked it constantly to prevent layering of the immiscible components of the charges.

When operating with high pressures in the cell, the oil in the chamber on the bottom side of the quartz crystal was automatically maintained at an equal pressure so that there was no tendency for the gaskets holding the crystal to leak. Furthermore, bending and damping of the crystal was minimized.

Since the reaction chamber was limited to charges of 200 cc or less, in order to treat large volumes of styrene at essentially atmospheric pressure, the cell was connected to a thermostatted ten-liter reservoir and a circulating diaphram pump by means of flexible metal hoses. The system was evacuated, the charge sucked in, and nitrogen forced in until the pressure was a few pounds per square inch above atmospheric. Samples of the latex could be removed at any time by means of a pet-cock.

The curves shown in Fig. 7 were obtained using the simple styrene formula given above in four-liter batches at 50°C with 200 watts plate input. As found with the magnetostriction oscillator, the main effect at this temperature was a decrease in the induction period and a somewhat higher final yield, although the actual rate of conversion was very little faster with the ultrasonic power on.

Figure 8 shows some results obtained at 40°C and, again, as found with the magnetostriction oscillator, there is a much more pronounced effect. Each of the three curves drawn in this plot was obtained with a different power concentration which was varied by changing the plate input to the oscillator or by changing the size of the batch.

No accurate method for measuring the actual ultrasonic power was available. Although a number of measurements were made on heating effects and radiation pressures, the figures obtained were unreliable. Probably about 50 percent of the plate input power was available as vibrational energy. However, to avoid the issue, the plate-input power was used as a basis for comparisons.

Figure 9 is a plot of the time required to reach a given yield of polystyrene as a function of the power concentration. These curves show that the reduction in time is approximately proportional to the power used. As the power is increased to very high concentrations, a minimum time to reach a given yield is approached. For

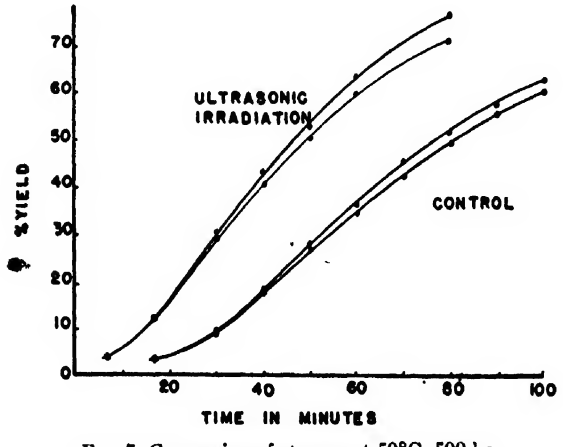


FIG. 7. Conversion of styrene at 50°C, 500 kc.

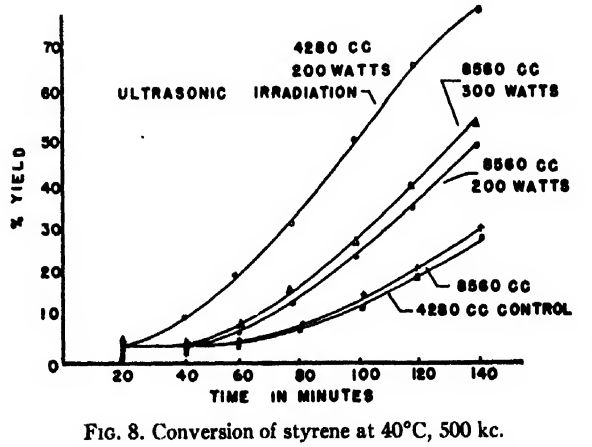


FIG. 8. Conversion of styrene at 40°C, 500 kc.

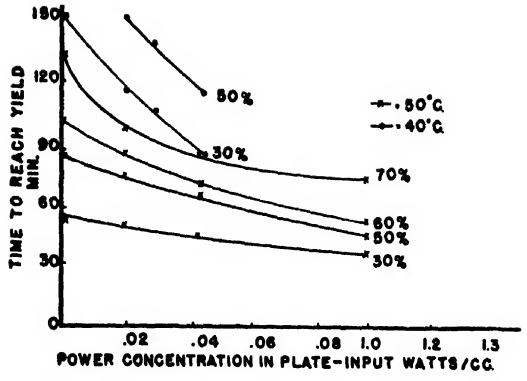


FIG. 9. Effect of power on conversion of styrene. (500-kc irradiation.)

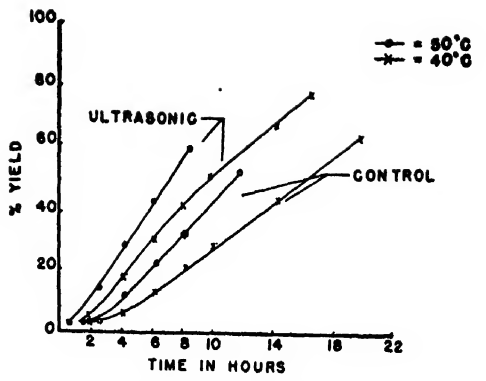


FIG. 10. Conversion of GR-S at 500 kc.

instance, at one watt per cc, 25 min. at 50°C is required to reach 70 percent yield.

As the power concentration is reduced to low levels by increasing the amount of latex in the batch, the curve intercepts the axis at the correct value as obtained from control runs with no power. However, if the batch is held at the same volume and the power concentration reduced by decreasing the plate input, a point is reached when no cavitation is produced. Below this critical intensity, the reaction rate is the same as the control, indicating that cavitation may be a necessary condition for the acceleration of the reaction.

IV. SNYTHETIC RUBBER POLYMERIZATION

For butadiene polymerization reactions, requiring 40 to 60 lb. pressure it was necessary to cut off the reservoir and circulating pump from the system, so that the batch size was limited to the reaction chamber. Figure 10 shows the results obtained for GR-S at 40° and 50°C. This was the standard mutual formula with a small amount of sulfuric acid added: 37.5 g of butadiene; 14.5 g of styrene; 90.0 g of distilled water; 2.5 g of soap; 0.15 g of potassium persulfate; 1.5 cc of 2.1*N* H_2SO_4; 1 to 2 drops of modifier (mercaptan).

The sulfuric acid solution was added to bring the *p*H down to 8.2 which was necessary to prevent foaming and consequent loss of ultrasonic power absorption by the latex. Control runs were made on the identical formula both in the reaction chamber without excitation of the crystal, and in standard glass bottles in a water bath. Samples were taken from the cell at definite intervals by means of a special hypodermic needle inserted through a gasket. The plate input was 200 watts, so that about 1 watt per cc (or about $\frac{1}{2}$ watt of vibrational power per cc) is required to double the yield in a given reaction time at either 40° or 50°C. This is a much larger power requirement than observed for the polymerization of styrene.

Essentially the same results were obtained with both a Redox "cold rubber" recipe at 41°F, as shown in Fig. 11, and a Redsol formula. The GR-S polymer produced with ultrasonic irradiation was apparently the same as the control product. Viscosity and density measurements showed no differences, and the infra-red absorption spectra were the same within the limits of spectrometer reproducibility.

V. DISCUSSION

The exact role of the ultrasonic agitation in accelerating these reactions is not understood. Apparently, cavitation is required, so that the high instantaneous pressures and temperatures produced upon collapse of the cavitation bubbles must be the principal cause of the effects. If higher localized temperatures at the cavitation centers are the cause, the polymer produced should have the characteristics of the control polymer made at higher temperatures. While no differences in the properties have been found as yet, the effects would be small and a more careful analysis of the properties may show these differences. If it were the high pressures (rather than temperatures) produced at the cavitation centers which were responsible for the rapid polymerizations, the properties of irradiated polymers should be compared to those special control polymers made at very high pressures. No such information is available at present.

In the case of the styrene polymerizations, the reduction of the induction period can be explained by the faster and better emulsification produced by the ultra-

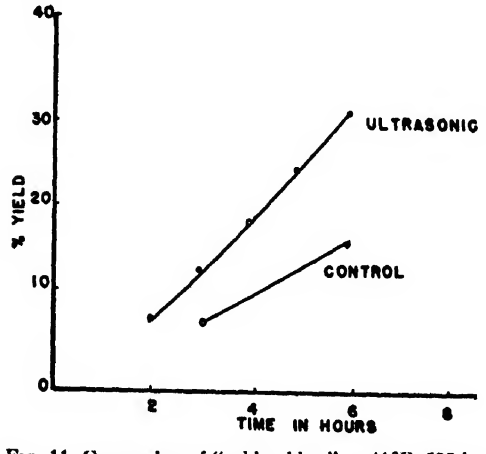


FIG. 11. Conversion of "cold rubber" at 41°F, 500 kc.

sonic agitation, a well-known phenomenon. In addition, several observers[3] have noted the ability of ultrasonic radiation to increase the rate of oxidation and produce peroxides. Since the induction period is usually caused by free oxygen in the system,[5] the rapid removal of this oxygen would permit the polymerization to start much sooner.

The increase in the actual rate of the reaction could be attributed to either increased rate of formation of free radicals by the persulfate reaction, or increased activity of the styrene in the polymer particles. The rate of reaction is usually considered to be established by the number of polymer particles formed initially. This is not greatly affected by irradiation (from particle size measurements) and the reaction cannot be "triggered" by the ultrasonic power. Several runs with both styrene and GR-S were made in which the ultrasonic power was cut off after a definite time. In these cases the reaction rate dropped to that of the control for the rest of the run.

The irradiated runs made with styrene usually proceeded to much higher yields (80 or 90 percent) before the reaction rate dropped off, while most control runs leveled off at 60 to 70 percent. This suggests that an important role of the ultrasonic radiation may be to oxidize inhibiting impurities which prevent the reaction from going to completion.

[5] I. M. Kolthoff and W. J. Dale, J. Am. Chem. Soc. 69, 441 (1947).

ACKNOWLEDGMENT

The authors wish to express their thanks to the Goodyear Tire and Rubber Company and to Dr. H. J. Osterhof for permission to publish this work. This investigation was carried out under the sponsorship of the Office of Rubber Reserve, Reconstruction Finance Corporation, in connection with the Government Synthetic Rubber Program.

Transitions in High Polymeric Materials

ROLF BUCHDAHL AND LAWRENCE E. NIELSEN

Monsanto Chemical Company, Plastics Research Laboratory, Springfield, Massachusetts

The evidence for and against the interpretation of the "second-order" transition in polymers in terms of a thermodynamic transition of the second order is reviewed and some new experimental data concerning transition in mixtures of polymers are presented. It is shown that the relaxation theory is insufficient to explain the transition, as observed by mechanical, electrical and thermal measurements. The usual objections which are raised against the assumption of a thermodynamic transition are shown to be invalid. Various molecular processes responsible for a thermodynamic transition are discussed and it is shown that the general mechanism of "hindered" rotation (changes in rotational configuration) is the most likely.

I

EMBRITTLEMENT has long been recognized as one of the characteristic phenomena of all polymeric materials and, from a phenomenological point of view, can best be described by the change of dynamic elastic modulus (E) and mechanical damping with temperature[1–4,*] as shown schematically in Fig. 1. Embrittlement occurs at a temperature, T_b, where the modulus has reached its maximum value. As the temperature is raised, the modulus undergoes a drastic change within a relatively narrow temperature range (the transition or "leathery" region) and then levels off again at higher temperatures where the material usually displays all the typical properties of an elastomer. In the transition the mechanical damping goes through a maximum, as can be seen from Fig. 1. Whether or not a material is classified at room temperature as a "plastic" or as a "rubber" depends only, from this point of view, on the position of the transition region on the temperature scale.

Two important characteristics of a damping temperature curve are the transition temperature defined as the temperature of maximum damping and the width. Within the transition region, as defined by dynamic mechanical measurements, many other physical properties change also in a more or less discontinuous manner. The specific heat undergoes a rather sudden increase as shown in Fig. 2; the thermal expansion coefficient[5] and the refractive index[6] changes, in most cases, in a discontinuous manner; the dielectric constant increases rapidly, and the electrical loss factor goes through a maximum.[7–9] (The dispersion in the electrical case is much more dependent on frequency than in the mechanical case.) Finally, in a few cases, it has been found

[1] A. W. Nolle, J. Polymer Sci. 5, 1 (1950).
[2] Nielsen, Buchdahl, Levreault, J. App. Phys. 31, 607 (1950).
[3] J. W. Liska, Ind. Eng. Chem. 36, 40 (1940).
[4] R. F. Clash, Jr. and R. M. Berg, Symposium on Plastics A.S.T.M. (1944).
* In references 3 and 4 a static modulus is used instead of the dynamic modulus.

[5] R. F. Boyer and R. S. Spencer, *Advances in Colloid Science II* (Interscience Publishers, Inc., New York, 1946), pp. 1–55.
[6] R. H. Wiley and G. M. Brauer, J. Polymer Sci. 3, 455 (1948).
[7] R. Fuoss, J. Am. Chem. Soc. 61, 2334 (1939); 63, 2401, 2410, 2832 (1941); 64, 283 (1942).
[8] Davies, Miller, and Busse, J. Am. Chem. Soc. 63, 361 (1941).
[9] Wurstlin, Kolloid Zeits. 105, 9 (1943).

that the compressibility[10] and thermal conductivity[11] undergo rapid changes in a narrow temperature region.

On the basis of the experimental evidence available, there can be no doubt that polymers undergo a transition which can be detected by numerous methods; there exists, however, a considerable amount of disagreement concerning the nature and origin of this transition. This problem can be stated quite simply as follows: (1) Is the transition a real one from the thermodynamic point of view? (2) If it is a real one, what is its origin? (3) If it is not real, in the thermodynamic sense, how can it be interpreted in a satisfactory manner? The fact that the transition has been referred to as a "second-order" transition indicates that the interpretation has, at times, been considered in terms of a real thermodynamic transition of the second order.[12] A thermodynamic second-order transition is defined as a transition extending over a small but finite temperature range ΔT with a heat of transition given by the integral $\int \Delta C dT$ where ΔC denotes the excess of specific heat at the transition.[13] However, during recent years the prevalent opinion has been that the transition is an apparent one and can be interpreted as the temperature "point where the rate of attainment of equilibrium is of the same order of magnitude as the time scale of ordinary physical experiments."[14] This conclusion is based on two separate but related experimental findings: The transition as measured by volume expansion shows strong time effects and when equilibrium conditions are approached tends to disappear (at least in the case of polystyrene);[15] the transition region shifts with frequency when measured by the response to an alternating mechanical or electrical force. According to this point of view, the transition is due to the existence of a relaxation spectrum[16] and corresponds to an isoviscous state.[17] Recently, however, several investigators[18-20] have emphasized the fact that the concept of relaxation is not sufficient to give a satisfactory interpretation of these transitions in high polymers and that it is necessary to assume some sort of thermodynamic transition. In the following sections we intend to show in a qualitative manner† (1) why such an assumption is necessary; (2) that the existence of a thermodynamic transition of the second order is in agreement with experimental findings and does not invalidate the existence of dispersion (or rate) effects; and finally we state the molecular process involved in the transition.

II

Before discussing the major limitations of any relaxation or isoviscous theory, it is necessary to consider the results Spencer and Boyer[20a] obtained in their study of the thermal expansion coefficient (α) of polystyrene. They concluded that, at least for polystyrene, the transition is a rate phenomenon and not a thermodynamic singularity. Polystyrene is a typical amorphous material where the mechanical properties are not further complicated by the presence of crystallites, and has only a single transition at about 105°C which is independent of molecular weight above $M_w \sim 150{,}000$;[21] it is important to clarify this point because it has frequently been used as the most powerful argument in favor of the relaxation theory. A careful examination of the published data does not seem to justify a linear relationship without a discontinuity between 20 and 150°C as claimed by the authors, unless the precision of the measurements is much lower than indicated. It seems that the discontinuity persists, irrespective of the time allowed to obtain equilibrium. It should be pointed out that the existence of a discontinuity is not a necessary criteria for the existence of a second-order transition; Fowler and Guggenheim have shown[22] that α above and below the transition temperature (T_t) is given by the

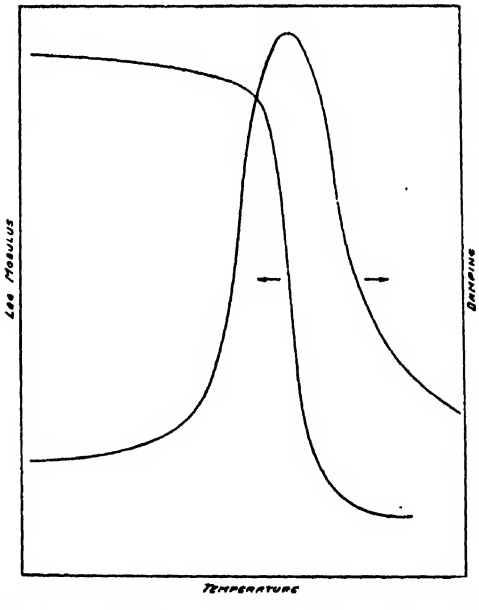


FIG. 1. Generalized dynamic modulus and damping curves as a function of temperature.

[10] A. H. Scott, J. Research Nat. Bur. Stand. 4, 99 (1935).
[11] A. Schallamach, Proc. Phys. Soc. London 53, 214 (1941).
[12] M. Ruhemann and F. Simon, Zeits. f. physik. Chemie A138, 1 (1928).
[13] J. Frenkel, *Kinetic Theory of Liquids* (Oxford University Press, New York, 1946).
[14] Alfrey, Goldfinger, and Mark, J. App. Phys. 14, 700 (1943).
[15] R. S. Spencer and R. F. Boyer, J. App. Phys. 17, 398 (1946).
[16] T. Alfrey, Jr., *Mechanical Behavior of High-Polymers* (Interscience Publishers, Inc., New York, 1946).
[17] R. F. Boyer and R. S. Spencer, J. App. Phys. 16, 594 (1945).
[18] Stein, Krimm, and Tobolsky, Phys. Rev. 75, 1301 (1949).
[19] E. Guth and H. M. James, Phys. Rev. 75, 1314 (1949).
[20] Merz, Nielsen, and Buchdahl, J. Polymer Sci. 4, 605 (1949).
† The quantitative formulation of such a theory will be published in the near future.

[20a] R. S. Spencer and R. F. Boyer, J. App. Phys. 17, 398 (1946).
[21] Merz, Buchdahl, and Nielsen, Ind. Eng. Chem. (to be published).
[22] *Statistical Thermodynamics*, Cambridge University Press, London (1939), p. 607.

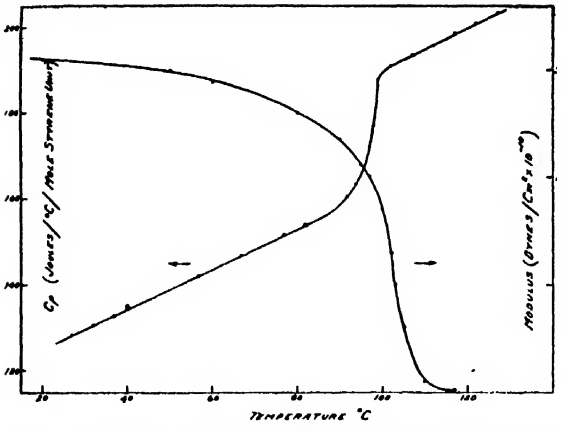


FIG. 2. Dynamic Young's modulus and specific heat as a function of temperature (the specific heat data were obtained by the N.B.S.).

following equations:

$$\alpha - \alpha_0 = -\beta^0(3/2)(w/kT_t^2)(dw/dv)(0 < T_t - T \leq T_t)$$
$$\alpha - \alpha_0 = 0(T > T_t),$$

where α_0 = expansion coefficient of the disordered state and w is the energy of unmixing. It will be noticed that a discontinuity in α at T_t only exists when dw/dv is different from zero or not negligibly small. There might well exist cases where the change in this energy with volume is sufficiently small or actually zero so that a discontinuity will not be found despite the existence of a thermodynamic transition.

One of the major difficulties encountered by the relaxation theory is the temperature dependance of the rate constant τ, near or in the transition region. It is generally assumed that τ varies exponentially with temperature: $\tau = C \cdot e^{W/RT}$, where W has the meaning of an activation energy. Such activation energies have been determined from measurements of electrical[23] or mechanical[1,24] damping, or from stress relaxation curves;[25] the values usually found are between 10 and 50 kcal. and appear to be closely related to the activation energies of viscous flow of high polymers.‡ However, these activation energies are much too low to account for the marked changes in the modulus in the transition region (as observed at low frequencies); in other words, the activation energy becomes itself a function of the temperature as the transition temperature is approached. As a matter of fact, Frenkel[13] has pointed out that the equation given above for τ must be modified by assuming W = const. eu/RT in order to be in agreement with experimental data; and that such an increase of the activation energy is due to a change of the structure which can be specified by an order-disorder parameter, s. It would seem more reasonable to postulate the following relationship for τ = const. $eW/[R(T-T_t)]$ for $T \geqq T_t$. It is therefore quite evident that the relaxation theory by itself cannot give a satisfactory interpretation of the strong temperature dependance of the modulus or damping, which in the case of unplasticized polyvinyl chloride is about 180 kcal., in terms of an activation energy, in the transition region.

Another difficulty arises when one considers the interrelationship between frequency (V) and temperature (T) with respect to the dynamic modulus and damping. On the basis of the relaxation theory there exists a definite interrelationship which can most easily be expressed by the proportionality

$$\ln V_1' \quad T_2$$
$$\ln V_2' \quad T_1'$$

i.e., a lowering of the temperature has the same effect as increasing the frequency, and $E(T)$ and $E(V)$ are closely related functions. Again, in the case of electrical measurements of simple low molecular weight materials in *non*-polar solvents, theory and experiment are in good agreement.[26] The extensive electrical studies by Fuoss[7] on various polymer-plasticizer systems also seem to be in harmony with a relaxation theory when a relaxation spectrum instead of a single relaxation time is assumed. It should be pointed out, however, that this agreement with theory might be somewhat fortuitous in view of the results obtained by Schallamach[27] on the dielectric relaxation of mixtures of dipolar liquids. When one now considers the available data[28,1] on mechanical damping or modulus as a function of frequency and temperature, it becomes immediately apparent that the agreement between theory and experiment is not good: (1) The frequency dependance is such that one has to assume a continuous relaxation spectrum extending over 10 decades or more; an assumption whose physical significance is—to say the least—somewhat doubtful. (2) An upper frequency limit of the modulus has never been experimentally established with certainty. (3) The interdependance between temperature and frequency is not very pronounced and certainly does not hold in the transition region. It will be shown in the next section that the dispersion effects can be interpreted as the consequence of the existence of a thermodynamic transition.

Finally, we wish to consider briefly the difficulties which arise with respect to the dependance of the transition temperature on plasticizer concentration on the basis of a relaxation or isoviscous state theory. ¶The

[23] W. Kaufmann, Rev. Mod. Phys. 14, 12 (1942).
[24] Alexandrov and Lazurkin, Acta Physicochimica (U.S.S.R.) 12, 647 (1940).
[25] A. Tobolsky and H. Eyring, J. Chem. Phys. 11, 125 (1943).
‡ On the basis of Debye's theory of relaxation $W_\tau = W_\eta$. Although this relation holds well for short and compact low molecular weight compounds in solution, it has been shown by Schallamach (Trans. Faraday Soc. 42, 495 (1946)) to be inapplicable in the case of non-polymeric chain molecules.

[26] P. P. Debye, *Polar Molecules* (Chemical Catalog Company, Inc., New York, 1929).
[27] A. Schallamach, Trans. Faraday Soc. 42A, 180 (1946).
[28] Sack, Lamb, and Work, J. App. Phys. 18, 450 (1947).
¶ This problem is intimately related to the general theory of plasticization which we plan to discuss in detail at a later date.

lowering of the transition point is explained in terms of a lowering of the internal viscosity so that upon addition of a plasticizer the isoviscous state (or the equivalent relaxation spectrum) has been shifted toward a lower temperature.[28a, 29] This concept implies that there can be only one transition point upon addition of a plasticizer. However, investigations of the dielectric losses in swollen rubber[30] show that one can obtain, under certain conditions, more than one maximum. Similar results have been found in the case of a mixture of two polymers (polystyrene and 40:60 percent butadiene-styrene copolymer) as can be seen in Fig. 3. Furthermore, a survey of reliable transition temperature data as a function of plasticizer concentration[2, 7, 8, 31] has shown that the transition temperature is a linear function of the volume fraction of plasticizer over a wide concentration range provided certain conditions are fulfilled[32, 2] (see Fig. 4). Recent investigations on the dielectric properties of plasticized compositions[33] have shown that the $T_c - c_v$ relationship when extrapolated to 100 percent plasticizers gives a transition point for the plasticizer which agrees with the experimentally determined transition point. The relationship derived by Boyer and Spencer[28a, ||] on the basis of the theory of isoviscous state is not in agreement with the $T_t - C_v$ relationship which, for the simplest case, is determined only by the transition points of the two components and has to be modified depending on the mutual compatibility of the constituents.

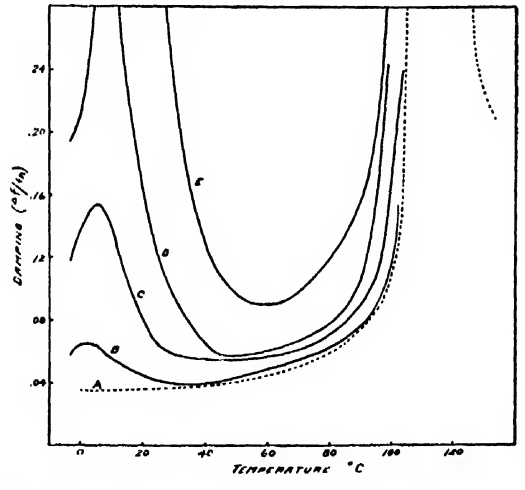


FIG. 3. Damping—temperature curves for polymer mixtures:

	Polystyrene	Styrene-butadiene copolymer
A	100	0
B	90	10
C	74	26
D	60	40
E	40	60

[28a] R. F. Boyer and S. Spencer, J. Polymer Sci. 2, 157 (1947).
[29] H. Jones, J. Soc. Chem. Ind. 67, 415 (1948).
[30] A. Schallamach, Trans. Faraday Soc. 45, 605 (1949).
[31] L. Nielsen and R. Levreault, Nature 164, 317 (1949).
[32] W. R. Richard and P. A. S. Smith, J. Chem. Phys. 18, 230 (1950).
[33] A. J. Warner, A.S.T.M. Bulletin (April, 1950).

|| In view of the fact that the Flory equation of viscous flow of polymers has recently been found to hold only over a limited molecular weight range, the agreement with experimental data is probably accidental.

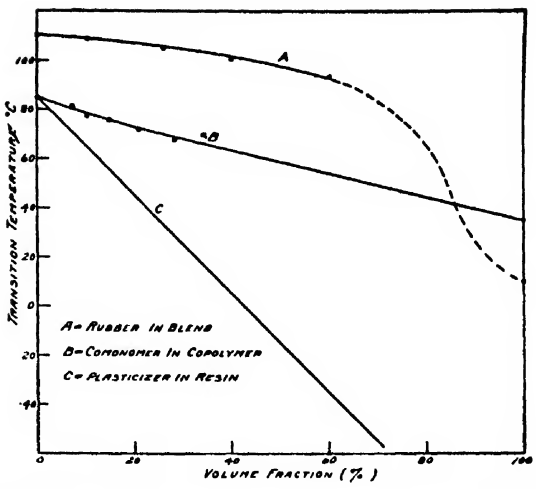


FIG. 4. Transition temperature as a function of volume fraction.

III

It is clear from the preceding discussion that the application of the relaxation concept to the transition phenomenon under consideration is insufficient to explain the salient features of such transitions. The necessity of assuming a real thermodynamic transition becomes, therefore, almost inescapable even when one approaches the problem at first from the relaxation point of view, as has been pointed out by several investigators.[1, 13] On the other hand, an interpretation on the basis of a thermodynamic transition has, in the past, encountered some difficulties because certain experimental findings appeared to be in disagreement with the concept of a true transition. These findings are (a) the apparent disappearance of the discontinuity of the thermal expansion coefficient in the case of polystyrene; (b) the shape of the specific heat curve in the transition region, and (c) the effect of time (or frequency) on the position and magnitude of the transition.

It has already been pointed out that the experimental data on equilibrium volume expansion of polystyrene confirm the existence of a discontinuity; in the one case where similar measurements were made on glass,[34] the discontinuity also persisted. Furthermore, recent measurements[35] of the volume expansion of various polystyrene fractions indicate that in some instances α goes through a maximum at the transition, similar to the specific heat changes observed in second-order transitions of simple low molecular weight compounds or metallic alloys.

At a second-order transition the specific heat changes in a characteristic manner: as the transition point is

[34] A. Winter, J. Am. Ceram. Soc. 26, 189 (1943).
[35] R. W. Raetz, Monsanto Chemical Company (unpublished data).

approached, it increases more and more rapidly; at the transition point, the specific heat falls back, in the case of small and simple compounds, to a value not much above the one it had many degrees below the transition temperature (λ-transition). Published data on the specific heats of polymers, such as the one shown in Fig. 2, are somewhat different; the specific heat increases rather abruptly in the transition region[36] but fails to return to a low value beyond the transition point. It has been argued[13,37,38] that specific heat curves of this type are characteristic of non-equilibrium conditions. This argument, however, overlooks the fact that, as a result of a second-order transition, additional degrees of freedom of the system might become operative, and the appearance of new degree of freedoms are not the cause but the result of the transition. As a consequence of this, the specific heat remains high and the discontinuity disappears, although very precise measurements over a very narrow temperature region within the transition range might or might not reveal a discontinuity; such data are not yet available for polymers. It should be pointed out that even in low molecular weight materials the transition is frequently not too sharp and that the diffuseness of the specific heat-temperature curve usually, but not necessarily, increases as the molecule becomes larger and more complex. Measurements on C_2F_6[39] illustrate this point very nicely and recent studies of hindered rotation on the same material[40] by means of the line width due to nuclear magnetic moments have shown conclusively that the specific heat change is due to a second-order transition.

Finally, there remains the effect of time or frequency on the position and magnitude of a second-order transition. Second-order transitions are usually free of relaxation effects, that is to say, the transition region or the transition temperature remain unaffected when the transition is studied by the application of external forces of different frequencies. In the case of polymers and glasses this is not so. Increasing the frequency shifts the transition temperature to higher values and also makes the transition region broader. This effect is easily understood when one remembers that the transition in polymers leads to a rubber, which on the basis of many extensive theoretical and experimental investigations[41–47] is the relaxing system par excellence; that is to say, its mechanical properties can quite satisfactorily be explained by a set or distribution of relaxation times. It is obvious, therefore, that as the material goes through the transition into the rubbery state, the relaxation effects will immediately become apparent. One should therefore expect that the transition is sharp when measurements are made at low frequency and becomes more diffuse for higher frequencies, because once the system has gone through the transition one deals with a rubber which will respond to forces of a given frequency according to its relaxation spectrum. This interpretation also explains the poor correlation between temperature and frequency in the transition region discussed in the previous section. The frequency dependance of the transition temperature is therefore not in conflict with the assumption of a true second-order transition but is due to the fact that the mechanical properties of one of the "phases" is determined by a widespread of relaxation times.

IV

It now remains to identify the thermodynamic transition with a reasonable molecular process. The absence of simultaneous electrical, mechanical and thermal data on one polymer makes it impossible to come to a definite conclusion on this point at the present time. However, there are a good many indications which point to one molecular process in preference to several others which *a priori* might be possible. These processes are: (1) melting of crystallites; (2) disappearance of long-range order as in the case of metallic alloys,[48] in ferromagnetic substances,[49] and possibly in some polar crystals;[13,50] (3) hindered rotation.[51,**]

The melting of crystallites can be ruled out as the cause of the transition for the following reasons. No latent heat of fusion has ever been found at the transitions under consideration, which means that no sharp discontinuity has been observed in the specific heat and the transition is not of the first order. It should be pointed out, however, that in the case of partly crystalline polymers, one does observe a transition in the mechanical properties;[52] this transition occurs though at considerably higher temperatures than the second-order transition point. Furthermore, in a study of the volume expansion coefficient (α) of amorphous and partly crystalline rubber, it has been shown[53] that the change in α of the latter material at the melting point is only a fraction of the change occurring in α at the second-order transition point, and that the change in α of the amorphous rubber is equal to the sum of the two

[36] Rands, Ferguson, and Prather, J. Research Nat. Bur. Stand. 33, 63 (1944).
[37] P. S. Ehrenfest, Comm. Leiden Supplement 75b (1933).
[38] E. Justi and M. Laue, Zeits. J. Tech. Physik 12, 521 (1934).
[39] E. L. Pace and J. G. Aston, J. Am. Chem. Soc. 70, 566 (1948).
[40] H. S. Gutowsky and G. E. Pake, J. Chem. Phys. 18, 162 (1950).
[41] P. Flory, Chem. Rev. 35, 51 (1944).
[42] H. M. James and E. Guth, J. App. Phys. 15, 294 (1944).
[43] M. S. Green and A. V. Tobolsky, J. Chem. Phys. 14, 80 (1946).
[44] G. Kirkwood, J. Chem. Phys. 14, 41 (1946).
[45] A. V. Tobolsky and R. D. Andrews, J. Chem. Phys. 13, 3 (1945).
[46] M. D. Stern and A. V. Tobolsky, J. Chem. Phys. 14, 93 (1946).
[47] Mooney, Wolstenholm, and Villars, J. App. Phys. 15, 324 (1944).
[48] F. C. Nix and W. Shockley, Rev. Mod. Phys. 10, 1 (1938).
[49] R. Becker and W. Doring, *Ferromagnetismus* (Edwards Brothers, New York, 1943).
[50] R. Frohlich, *Theory of Dielectrics* (Oxford University Press, New York, 1949).
[51] L. Pauling, Phys. Rev. 36, 430 (1930).
** The term "hindered rotation" is used here in the broadest possible sense, including, for example, *cis-trans*-isomerism.
[52] L. Nielsen, Monsanto Chemical Company (unpublished data).
[53] N. Bekkedahl, J. Research Nat. Bur. Stand. 13, 411 (1934).

separate changes in the partly crystalline material. In other words, the fraction of material which has entered into a crystalline structure does not contribute significantly to the second-order transition at lower temperature. No noticeable change in the x-ray diffraction pattern has been observed in the transition region.

The disappearance of long-range order as a cause of the transition is most unlikely because a transition of that kind presupposes the existence of a "superlattice" for which there exists no experimental evidence in high polymers, at the present time. The rather marked change in the volume expansion coefficient at the transition is another argument against long-range-short-range order transition. Finally, the frequency dependance of the transition—the fact that the transition becomes sharper at lower frequencies and appears to approach a limiting value as $V \to 0$—argues against a transition based primarily on cooperative phenomena.[19]

This leaves the process of hindered rotation[22, 51, 54]—applied here in a very general sense—as the cause of the second-order transition. Experimental evidence available so far and analogies with similar transitions in low molecular weight compounds all favor an interpretation on this basis. First of all, we should recall the volume expansion data on amorphous and partially crystalline rubber, referred to. The non-crystalline phase has its transition at the same temperature and the change in α is of the same type in the amorphous region as it is in the crystalline region. At a given temperature there occurs an "internal" or molecular freezing.[55] This phenomenon has been observed repeatedly in a great many low molecular weight compounds.[54, 56] Such transitions, of which there can be more than one in one compound, are a characteristic property of all solid matter. The close similarity between low and high molecular weight materials is brought out very clearly by the concentration dependance of the transition temperature in polymer-plasticizer compositions.[7, 8, 31, 33, 57] It has already been stated that for several plasticizers transition points have been found experimentally which agree well with extrapolated values from the linear portion of the $T_t - C_v$ curves. It is fortunate that the dielectric and thermal properties of one of these plasticizers, camphor, has been studied extensively[58, 59] where it has been shown that the transitions of this and similar compounds is definitely due to hindered rotation. These studies on camphor furthermore reveal how strongly the transition temperature and the shape of the dielectric constant-temperature curve is affected by slight differences in molecular structure such as *cis-trans*-isomerism. The same has been observed in polymers. Unvulcanized rubber and gutta percha are considered *cis-trans*-isomers exhibiting brittle points at $-58°C$ and $-23°C$, respectively; similarly, the isomers discovered by Schildknecht[60] show characteristic differences in their mechanical properties at room temperature which is indicative of different transition temperatures. It is also of interest to point out that vulcanization leads to an increase in the transition temperature.[5] This has recently been interpreted by Nolle[1] as due to a decrease of the effective chain length between "hindrances." Although vulcanization leads undoubtedly to a large number of crosslinks, it is not the reduced chain lengths which increases the transition temperature, but the increased hindrances which are introduced by sufficiently large amounts of crosslinking.

It appears, therefore, that all available data can be satisfactorily interpreted on the basis of a second-order transition which is due to hindered rotation of complex molecules. The detailed knowledge of the molecular potential which will permit the establishment of a reliable qualitative theory will require more extensive investigation of mechanical, electrical, and thermal properties. The recently developed method of magnetic nuclear resonance[40, 61] should be particularly beneficial.

[54] C. P. Smyth, Chem. Rev. **19**, 329 (1936).
[55] K. Uberreiter, Zeits. f. angew Chemie **53**, 247 (1940).
[56] A. Eucken, Zeits. f. Elektrochemie **45**, 126 (1939).
[57] M. C. Reed and J. Harding, Ind. Eng. Chem. **41**, 675 (1949).
[58] W. A. Yager and S. O. Morgan, J. Am. Chem. Soc. **57**, 2071 (1935).
[59] A. H. White and S. O. Morgan, J. Am. Chem. Soc. **57**, 2078 (1935).
[60] Schildknecht, Gross, and Zoss, Ind. Eng. Chem. **41**, 1998 (1949).
[61] Gutowsky, Kistiakowsky, Pake, and Purcell, J. Chem. Phys. **17**, 972 (1949).

Mechanical Properties of Oriented Polystyrene Film

LAWRENCE E. NIELSEN AND ROLF BUCHDAHL
Monsanto Chemical Company, Plastics Research Laboratory, Springfield, Massachusetts

Cast films of polystyrene were oriented by stretching at temperatures above the softening temperature of the material and then cooled. A factorially designed experiment was carried out in which the amount of stretching, temperature of stretching, time held in the stretched state before cooling, and effect of several cycles of stretching before cooling were investigated as they affect the birefringence, the stress-strain properties, dynamic modulus and damping, and the softening temperature of polystyrene.

The values of many of the mechanical properties are more closely related to the birefringence than to the amount of hot stretching. In general, the value of the stress-strain properties in the direction of hot stretch, the dynamic modulus, and the birefringence all increase with an increase in the amount of stretching, a decrease in temperature, an increase in rate of cooling after stretching, and a decrease in the number of times the film was stretched before finally cooling. The softening temperature is decreased in direct proportion to the birefringence. Annealed films give higher softening temperatures than films which have been stretched and then allowed to shrink before cooling.

IT is quite well known that often remarkable changes in the physical properties of a high polymeric substance may be brought about by stretching the material while it is at a temperature above its softening point and then cooling it down.[1-3] However, practically no quantitative data have been published showing how the mechanical properties such as the stress-strain characteristics are changed by such a treatment.

This stretching process undoubtedly causes the high polymer molecules to become oriented or arranged in a configuration different from their normal arrangement. On cooling after stretching, the material changes from a soft rubber-like state to a rigid state, thereby "freezing" the molecules in an oriented condition. These lined up molecules, which have different characteristics parallel and perpendicular to their long dimension, give an anisotropic behavior to many physical properties. For instance, the stretched material may show optical anisotropy or birefringence, and the mechanical properties such as ultimate strength may differ greatly parallel and perpendicular to the direction of stretch.

Orientation phenomena are very complex in nature not only because there are many variables such as the amount of stretch and the temperature at which stretching took place but also because of the fact that there are different kinds of orientation. In this investigation only stretching in one direction will be considered. Furthermore, all the data and discussion to follow except on birefringence apply only to the direction of stretch and not to the direction perpendicular to it. It is to be expected that entirely different properties will be found in the direction perpendicular to the stretch.

The problem of quantitatively measuring the degree of orientation arises. On the basis of past work it appears that the birefringence is one of the best methods of measuring the orientation. The amount of hot stretching is not satisfactory because of the tendency of the molecular segments to assume an unoriented condition. Therefore, for a given amount of hot stretching the number of oriented molecular segments found in the cooled sample depends upon the temperature of stretching, how fast the specimen was stretched, and how long it took the specimen to cool down.

All oriented specimens were made up according to the factorial design given in Table I. A factorially designed experiment was chosen because several variables were to be studied and it was desirable to know to what extent these variables interact with one another. In addition it is possible to obtain the largest amount of

TABLE I. Design of experiment on oriented polystyrene films.

Elongation percent	50				150				300			
After treatment	Chilled		Relaxed		Chilled		Relaxed		Chilled		Relaxed	
Amount of working	O	X	O	X	O	X	O	X	O	X	O	X
Temperature	105	105	105	105	105	105	105	105	105	105	105	105
	115	115	115	115	115	115	115	115	115	115	115	115
Specimen No. 1	1CO11	1CX11	1RO11	1RX11	2CO11	2CX11	2RO11	2RX11	3CO11	3CX11	3RO11	3RX11
	1CO21	1CX21	1RO21	1RX21	2CO21	2CX21	2RO21	2RX21	3CO21	3CX21	3RO21	3RX21
Specimen No. 2	1CO12	1CX12	1RO12	1RX12	2CO12	2CX12	2RO12	2RX12	3CO12	3CX12	3RO12	3RX12
	1CO22	1CX22	1RO22	1RX22	2CO22	2CX22	2RO22	2RX22	3CO22	3CX22	3RO22	3RX22

[1] J. Bailey, India Rubber World **118**, 225 (May, 1948).
[2] H. Müller, Kolloid Zeits. **95**, 138, 306 (1941).
[3] H. Müller, Wissenschaftliche Veröffentlichungen aus den Siemens Werken **19**, 110 (1940).

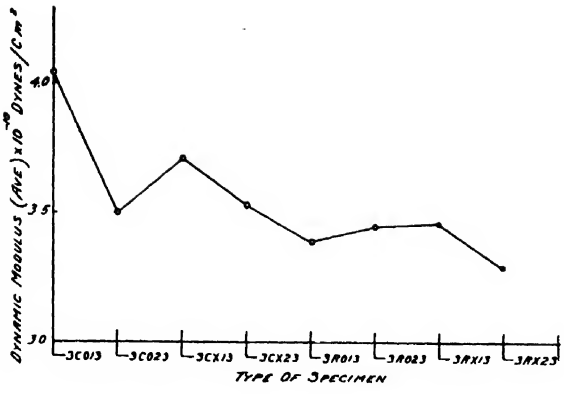


FIG. 1. Dynamic Young's modulus of oriented polystyrene films at 25°C.

information with the smallest amount of work with this type of experiment. All specimens were made from strips of film which before stretching were one inch wide and roughly 0.015-inch thick. These films had been cast from methyl ethyl ketone and were carefully dried and annealed for several days as previously described.[4] Studies were made at three levels of elongation—50 percent 150 percent, and 300 percent hot stretch. Two types of after-treatment were used. In the one case the film was immediately chilled to below the softening temperature after stretching. In the other case the film was held at its final length for 200 seconds before cooling. It has been shown[5] that the birefringence and retractive force of the film are greatly decreased from their initial value after 200 seconds, even though the length has been held constant. Therefore, it was important to be able to separate effects due to elongation from those determined primarily by birefringence or "frozen in" stresses. From our previous work,[5] it had been found that such properties as creep depend upon whether the film had been used previously for such a test. Therefore, for lack of a better term the effect of "working" was investigated. In Table I "O" stands for stretching with no previous working while X means the sample was twice stretched and then allowed to shrink to nearly its original length before the film was stretched again and then cooled. Stretching was carried out at two different temperature levels. At 105°C the film is quite stiff and only slightly rubber-like in nature while at 115°C the polystyrene films are very much like a rubber. In the code numbers to designate the specimens the first numbers 1, 2, and 3 stand for 50 percent, 150 percent, and 300 percent elongation, respectively. Figures 1 and 2 in fourth position of the code numbers indicate 105°C and 115°C, respectively. The number of the sample is given by the last number of the code. For example, a code number 2CO13 would mean the sample had been elongated 150 percent and immediately chilled without working at a temperature of 105°C and that it was the third such specimen prepared. In addition to these 24 types of oriented films some experiments for comparison were made on annealed unoriented films and worked unoriented films. These latter were prepared by stretching and shrinking the film twice before cooling down with no stretching. Actually in most cases the worked films did not completely shrink to their original length; the birefringence, though, was either zero or very small compared to the oriented films.

The following measurements were made on these films: birefringence, tensile stress-strain tests, softening temperature, dynamic Young's modulus, and mechanical damping.

DYNAMIC MECHANICAL TESTS

Since it was suspected that the dynamic Young's modulus and mechanical damping were not radically changed by orientation in the case of polystyrene, only specimens elongated 300 percent were studied. The dynamic properties were determined with an electromagnetic reed vibrator.[4, 6–8] The modulus was calculated from the resonance frequency and the damping from the half-width of the resonance peak. (The half-width of the resonance peak may be expressed as the ratio of the imaginary to real part of Young's modulus by dividing the half-width by $\sqrt{3}$.) The results at 25°C and 30 c.p.s. are summarized in Table II and Fig. 1. Results on annealed and worked films are also included.

Figure 2 compares the damping (half-width of resonance peak) for annealed, worked, and typical oriented (3RX23) films as the temperature is varied.

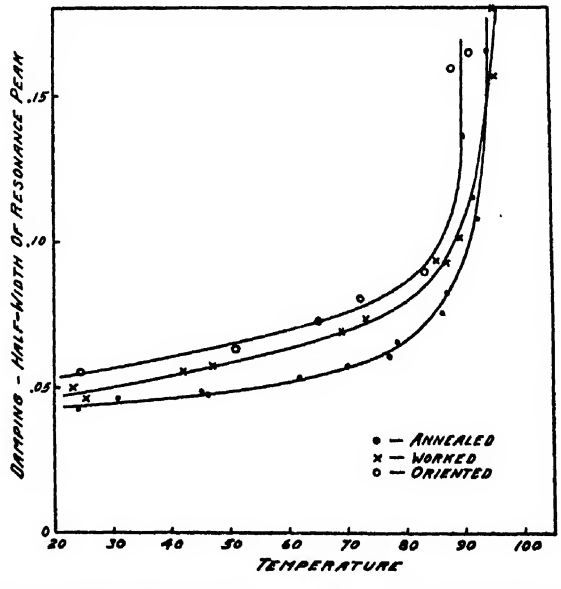


FIG. 2. Mechanical damping as a function of temperature for annealed, worked, and typical oriented polystyrene films.

[4] Merz, Nielsen, and Buchdahl, J. Polymer Sci. 4, 605 (1949).
[5] L. E. Nielsen and R. Buchdahl, J. Chem. Phys. 17, 839 (1949) and J. Colloid Sci. (to be published).
[6] L. E. Nielsen, A.S.T.M. Bulletin 165, 48 (1950).
[7] A. W. Nolle, J. App. Phys. 19, 753 (1948).
[8] J. Ballou and J. Smith, J. App. Phys. 20, 493 (1949).

TABLE II. Dynamic mechanical properties of oriented and unoriented polystyrene films at 25°C.

Type of film	Birefringence A/mil (av.)	Dynamic Young's modulus (dynes/cm²)	Half-width of resonance peak	No. of specimens tested
Annealed	0	$3.24\pm0.10\times10^{10}$	0.046 ± 0.002	5
Worked	~0	$3.18\pm0.14\times10^{10}$	0.049 ± 0.0015	4
3CO13	4500	$4.04\pm0.20\times10^{10}$	0.054 ± 0.002	6
3CO23	2640	$3.50\pm0.09\times10^{10}$	0.049 ± 0.001	3
3CX13	4280	$3.71\pm0.19\times10^{10}$	0.053 ± 0.004	3
3CX23	1970	$3.53\pm0.24\times10^{10}$	0.051 ± 0.002	4
3RO13	2090	$3.39\pm0.08\times10^{10}$	0.051 ± 0.002	3
3RO23	864	$3.45\pm0.13\times10^{10}$	0.056 ± 0.004	3
3RX13	1580	$3.46\pm0.12\times10^{10}$	0.050 ± 0.004	3
3RX23	814	$3.29\pm0.16\times10^{10}$	0.056 ± 0.000	3

On the basis of these results, it is seen that orientation increases the modulus of polystyrene. Comparison with birefringence data indicates the dynamic modulus varies in about the same way as the birefringence. Damping also increases with orientation but only slightly and part of this increase may be attributed to the effect of working which automatically takes place during the process of stretching the film. Both orientation and working lower the transition temperature range where the damping increases very rapidly with temperature.

BIREFRINGENCE AND STRESS-STRAIN DATA

The birefringence and stress-strain data are given in Table III. The notation is the same as used in Table I. Duplicates were made of all the specimens.

The birefringence was measured by a Babinet compensator using polarized light. Figure 3 indicates how the average birefringence changes with the different types of specimens. The birefringence of films stretched at 105°C is always considerably higher than that of films stretched at 115°C, the other conditions being held constant. Chilling the film immediately after stretching gives a higher birefringence than if the film is held in the stretched state for 200 seconds before cooling. Of course, the greater the elongation the greater is the birefringence if the other variables are held constant. Working has very little effect. An analysis of variance[9] was carried out on these data and the effects of the variables temperature, amount of stretching, and after-treatment were found to be significant at the 99.9 percent level of confidence. There seems to be no interaction of the variables in the case of birefringence.

The stress-strain data were obtained on an Instron machine[10] using a crosshead speed of 0.2 inch/minute and a specimen length of 2 inches between the clamps. The tests were conducted at a temperature of 25°C under a relative humidity of 50 percent. Table III also contains data on unoriented annealed and unoriented worked films for comparison; these data are the average

TABLE III. Birefringence and stress-strain data on oriented polystyrene films.

Specimen	Birefringence (A per mil $\times10^{-2}$)	Modulus (psi $\times10^{-5}$)	Elongation at yield (percent)	Elongation to break (percent)	Yield stress (psi)	Ultimate strength (psi)	Energy to break (arbitrary units
1CO11	11.2	2.68	7.40	7.40	10480	10480	3670
1CO12	13.2	3.08	—	5.80	—	10420	4790
1CO21	3.35	2.88	4.75	4.75	7600	7600	2090
1CO22	4.72	2.49	4.65	4.65	8560	8560	2130
1CX11	9.03	3.72	—	5.00	—	9580	2660
1CX12	11.3	2.85	—	4.90	—	9620	2730
1CX21	3.20	2.17	—	3.70	—	7850	1610
1CX22	5.07	2.79	—	4.25	—	7870	1990
1RO11	2.64	2.73	—	3.85	—	7700	1740
1RO12	3.91	2.79	—	4.00	—	7730	1760
1RO21	1.24	3.03	3.25	4.00	7130	6850	1880
1RO22	2.01	3.46	—	3.60	—	7300	1510
1RX11	2.45	2.91	—	3.90	—	7600	1650
1RX12	3.29	3.03	—	3.70	—	7460	1550
1RX21	1.18	2.97	3.25	3.50	6680	6620	1360
1RX22	1.86	2.70	4.40	4.40	6440	6440	1750
2CO11	32.2	4.17	4.65	4.90	10850	10650	3620
2CO12	28.4	3.45	5.55	6.85	10820	8870	5280
2CO21	10.7	3.52	5.20	5.30	10400	10400	3460
2CO22	12.8	3.51	5.25	5.60	10100	9380	3920
2CX11	24.0	3.74	5.30	5.60	11250	11250	4000
2CX12	26.8	3.68	4.75	5.00	10380	9820	3240
2CX21	9.22	3.52	—	4.00	—	9120	2180
2CX22	12.2	3.40	4.85	5.50	10900	9320	3670
2RO11	8.42	3.33	5.10	5.10	10250	10250	3310
2RO12	9.69	3.25	5.10	5.10	10240	10240	3310
2RO21	4.48	3.11	4.00	4.90	8770	8550	2290
2RO22	5.16	3.11	—	3.85	—	8940	2150
2RX11	10.0	3.49	—	4.00	—	9350	2170
2RX12	8.68	3.42	—	4.25	—	10180	2390
2RX21	4.95	3.22	4.20	4.20	8590	8590	2170
2RX22	4.89	3.15	—	3.45	—	8740	1630
3CO11	50.6	3.74	4.60	6.35	10100	9540	4720
3CO12	65.8	4.81	4.50	6.20	10810	10000	2170
3CO21	41.2	4.64	4.35	5.35	11770	9700	4380
3CO22	31.4	3.75	4.75	7.50	10350	8290	5730
3CX11	47.0	4.25	4.40	5.70	10950	10420	4470
3CX12	35.7	3.63	4.25	4.95	10340	9740	3450
3CX21	22.3	3.68	4.40	5.05	9730	9170	3460
3CX22	33.2	3.51	4.25	5.05	10560	9340	3590
3RO11	23.4	3.46	4.50	5.75	10300	9160	4080
3RO12	18.5	3.61	4.60	9.25	10780	7330	6960
3RO21	11.0	3.71	4.30	4.60	10360	10100	3220
3RO22	10.9	3.70	4.25	4.25	10640	10640	2900
3RX11	17.9	2.98	4.25	4.75	9310	8660	2770
3RX12	29.7	3.63	4.25	4.70	10570	8850	3270
3RX21	10.3	3.22	—	4.20	—	9800	2290
3RX22	8.55	3.66	4.05	4.05	10000	10000	2530
Annealed (av.)	0	2.66	—	2.53	—	6170	
"Worked" (av.)	0.35	2.70	3.00	3.00	6720	6720	

of 3 specimens each. The blank spaces in the columns for yield data arise from the fact that not all of the

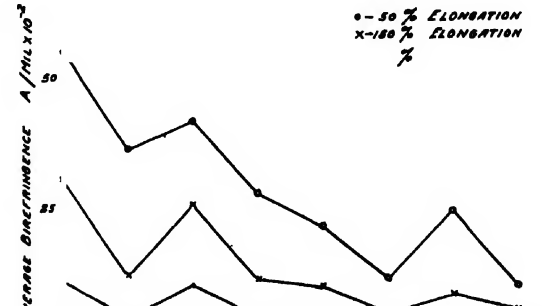


FIG. 3. Birefringence of oriented polystyrene films prepared under various conditions. The types of specimen are designated by the code given in Table I.

[9] K. A. Brownlee, *Industrial Experimentation* (Chemical Publishing Company, Inc., Brooklyn, New York, 1947).
[10] G. S. Burr, Electronics 22, No. 5, 101 (1949).

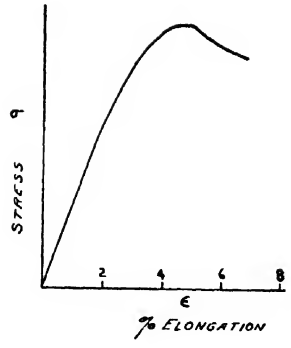


FIG. 4. A typical stress-strain curve for oriented polystyrene films.

specimens gave a yield value. A typical stress-strain curve is given in Fig. 4.

The variation in Young's modulus for the different types of samples is illustrated in Fig. 5. It will be noted that the modulus shows some of the same general trends as the birefringence. The modulus increases with the amount of hot stretching and usually decreases with temperature. However, the analysis of variance indicates the effects of these variables to be at only about a 90 percent level of significance and are largely covered up by a large interaction between elongation and after-treatment. The effect of working and after-treatment on the modulus is not significant. All the oriented films have higher moduli than annealed or unoriented worked films.

The analysis of the data on elongation to break the specimens indicates the elongation increases with the amount of stretching (99 percent level of significance), it decreases with increase in temperature (99 percent level of significance) and it decreases with working (99.9 percent level of confidence). The elongation to break is greater for chilled than relaxed specimens with a level of confidence of 99.9 percent. There are no interactions between the variables for the data on elongation to break. During the tests it was noted that if a yield point was reached, all further elongation took place near this yield point in a small band across the specimen. Much larger elongations would have been observed if all the specimen had elongated as much as that portion near the original point of yielding.

The energy to break a specimen, that is, the area under the stress-strain curve, is relatively free of interactions between the variables also. The more the specimen is stretched while soft the greater is the energy to break (99.9 percent level of significance). Increasing the temperature decreases the area under the stress-strain curve (99 percent level of significance). The values are larger for chilled than relaxed specimens (99.9 percent level of significance). Working decreases the energy to break (99 percent level of significance). The energy to break also shows the same general trends as the birefringence.

The ultimate strength is a complex function of the variables studied, although all the oriented specimens gave higher values than the unoriented ones. The analysis of variance indicated the most significant factor was a very strong interaction between the three variables, amount of stretch, after-treatment, and temperature. Undoubtedly, some of this complexity is brought about by the fact that some of the specimens had yield points and others did not. Most of the specimens which did show yield points gave low values of tensile stress once the yield point was passed, as is shown in Fig. 4. Those films with the greatest amount of hot stretching in general gave yield values while the films with the least amount of stretching broke at the yield point or before. Therefore, the expected increase in ultimate strength with amount of stretching was not picked up.

An analysis of variance was not made on the elongation at yield and the stress at the yield point since some of the films did not show a yield point. Certain general conclusions may be made though. The greater the amount of hot stretching the greater is the possibility of reaching a yield point. Most of the films with 50 percent stretching broke at the yield point or before it was reached. Neither the elongation at yield nor the yield stress depends to any great extent on the temperature of stretching, after-treatment, or working. Although none of the unoriented annealed films gave a yield point, all the unoriented worked films did, surprisingly enough.

In the cases where there are no interactions of the variables as shown by the analysis of variance, it is possible to calculate quite accurately the variation of the properties as the variables are changed. These results are summarized in Table IV.

Since many of the mechanical properties appear to vary in the same manner as the birefringence, correlation coefficients were determined. The results are summarized in Table V. These correlation coefficients were found for the following equations:

$$10^{-5}E = 2.50 + 0.85 \log B$$
$$\epsilon_b = 3.26 + 1.65 \log B$$
$$\sigma_b = 5708 + 3510 \log \log B$$
$$\log(A - 3.00) = 0.13 + 0.32 \log B.$$

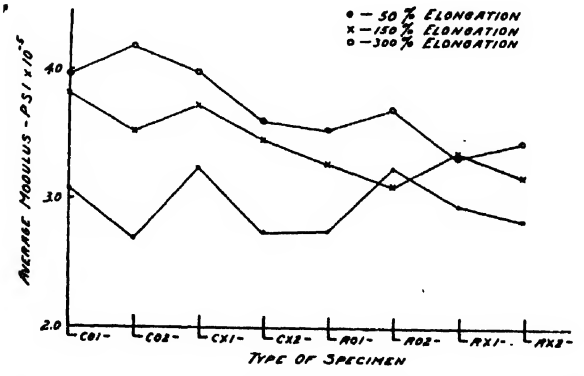


FIG. 5. Young's modulus calculated from the initial slope of the stress-strain curves of oriented polystyrene films. The code designating the types of specimens is given in Table I.

TABLE IV. Relative value of properties on changes of variables.[a]

Property	ϵ_2/ϵ_1	ϵ_3/ϵ_1	T_2/T_1	t_2/t_1	W_2/W_1
Birefringence	2.76	5.74	0.52	0.39	1.0
Elong. @ break	1.09	1.23	0.86	0.83	0.84
Energy to break	1.40	1.72	0.80	0.73	0.77

[a] The reference values are: ϵ_1 = 50 percent hot stretch; T_1 = 105°C; t_1 = chilling after stretching; W_1 = no working. The other variables are: ϵ_2 = 150 percent; ϵ_3 = 300 percent; T_2 = 115°C; t_2 = relaxed before cooling; W_2 = working.

TABLE V. Correlation of birefringence with stress-strain properties.

Variables	Correlation coefficient
Birefringence and Young's modulus	0.813
Birefringence and elongation	0.717
Birefringence and tensile strength	0.802
Birefringence and energy to break	0.864

In these equations B = birefringence in A/mil$\times 10^{-2}$, E = Young's modulus in pounds/square inch, ϵ_b = percent elongation at break, σ_b = tensile strength in pounds/sq. in., and A = the energy to break in arbitrary units.

Table V shows that the stress-strain properties of oriented polystyrene films are very closely related to their birefringence. It is known that birefringence is a fairly good quantitative measure of orientation, and it appears that the mechanical properties are primarily dependent also upon this degree of orientation. The effects of such variables as temperature, after-treatment, and working are, to a first approximation, important only to the extent they change the birefringence. However, the importance of these other variables cannot be completely overlooked in interpreting the stress-strain properties of oriented polystyrene.

SOFTENING TEMPERATURES

One of the most important properties of plastics from the practical standpoint is a softening temperature. This softening temperature may be defined in various ways, but in all cases it gives an indication of the temperature range where the material goes through a transition from a hard, rigid substance to a soft, rubbery one.

If an oriented film is placed in an oven in which the temperature is gradually increasing, the length of film will remain nearly constant until the softening temperature is reached. Over a very narrow temperature range, the film will then rapidly shrink in length as the orientation starts to disappear. Müller[3] has shown that the softening temperature is lowered as the birefringence of the oriented films increases. This has been checked using films whose birefringence has been varied by different amounts of stretch, stretching at different temperatures, and by different after-treatments. Results are given in Table VI and Fig. 6. The softening temperature was arbitrarily defined as that temperature at

TABLE VI. Change in softening temperature with birefringence.

Film No.	Birefringence A/mil.	Softening temperature °C
2RX23	400	95.3
2CO23	1020	93.5
2CX13	2150	91.0
3CO13	5950	83.2

which the specimen had shrunk one percent. The temperature of the oven was increased at the rate of about ½°C per minute.

From these results, it is obvious that the lowering of the softening temperature is primarily dependent upon the birefringence and not upon other possible variables such as elongation or after-treatment.

The softening temperature is also altered by an externally applied force[11] or by working. The effect of stress was verified by gradually heating the specimens to which tensile loads of different sizes were applied by hanging weights from clamps. The length of specimen remains nearly constant till a very sudden increase in length takes place on reaching the softening range. Some of the data are illustrated in Fig. 7, where the softening temperature was again taken as the temperature where the total elongation was 1 percent. The worked films give a softening temperature several degrees lower than annealed films. Since some working automatically takes place in the process of orienting the films by stretching, it is to be expected that the softening temperature extrapolated to zero birefringence (see Fig. 6) would be lower than the softening temperature obtained by extrapolating to zero external force (see Fig. 7). This turns out to be the case, the values being 96.6°C and 98.0°C respectively.

DISCUSSION

On the basis of the data presented here and in the past,[3,6] it has been shown that orientation or the

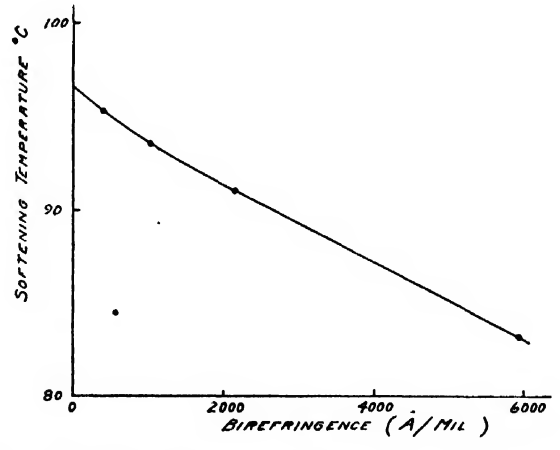


FIG. 6. Variation of softening temperature with the birefringence of oriented films.

[11] R. F. Boyer and R. S. Spencer, J. App. Phys. 16, 594 (1945).

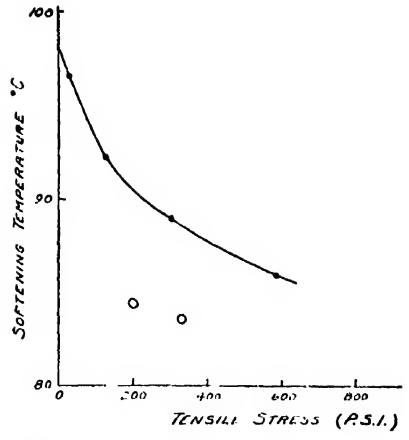


FIG. 7. Variation of softening temperature with tensile load for annealed polystyrene films. The two points enclosed in circles are for unoriented worked films.

lining up of molecular segments is far more accurately measured by birefringence than by the amount of hot stretching in the case of polystyrene. Widely different values of birefringence may be obtained with the same amount of stretching by changing the temperature at which stretching takes place, or by holding the hot film in the stretched state for some time to allow some of the molecular segments to resume their normal relaxed state. Such properties as the softening temperature, and elastic modulus reflect the changes in birefringence which are brought about by these different treatments.

We attribute the decreased birefringence and changes in mechanical properties which accompany an increase in the stretching temperature as due primarily to the molecules being able to assume a relaxed or unoriented condition more rapidly as the temperature increases. Similarly, most of the changes brought about by the after-treatments, such as chilling immediately after stretching or allowing some time to elapse before cooling, may be explained in terms of the degree of orientation "frozen" into the cooled films.

Some of the phenomena which appear to be influenced by "working" may be due to the partial destruction of the network structure of entangled and intertwined molecules found in polystyrene.[5] Other effects such as the lowering of the softening temperature are believed to be due to non-equilibrium conditions existing in worked films. In an annealed film the molecules have had time to assume their most probable shapes and configurations. However, on stretching and shrinking a film, many molecular segments are forced into positions and shapes not most favorable for them; this working would then be expected to decrease the density of the material over that found under equilibrium conditions existing in annealed films. This slight increase in the average distance between molecules of the worked films would be expected to lower the softening temperature somewhat.

We wish to express our appreciation to Mr. C. H. Adams and to the Fabrics Research Laboratory for their help in securing the stress-strain data and to Mr. E. C. Harrington who carried out the statistical analysis of the data.

A Non-Destructive Mechanical Test for Animal Fibers

HARRIS M. BURTE*
The Textile Research Institute and the Department of Chemical Engineering, Princeton University, Princeton, New Jersey

The slope of the force-extension curve in the Hooke's law region can be measured non-destructively, for animal fibers, at three to fifteen minute intervals. The change of this Hookean slope with time is used to follow the course of the heterogeneous reaction between wool and a reagent.

Data are presented on the interaction between concentrated neutral salt solutions and wool fibers. These results indicate that wo processes occur: (a) Relatively rapid hydration or dehydration of the fiber until equilibrium with the activity of water in the salt solution is attained; (b) Relatively slow absorption of salt ions by the fiber. Very concentrated salt solutions dehydrate the fiber so completely that absorption of salt ions is not possible. Swelling experiments and experiments involving other mechanical properties confirm these hypotheses. The interaction between wool fibers and aqueous solutions of large organic molecules follows a similar pattern.

MOST natural high polymers are chemically and morphologically quite complex. A study of the mechanisms and kinetics of their reactions using the classical methods of chemical analysis is frequently a laborious task. Their mechanical properties, however, are easily measured, and usually are strongly affected by chemical treatment. The investigation of chemical reactions by using a mechanical property as the measured parameter has therefore attained some popularity. The present work is concerned with a further extension of these mechano-chemical methods[1] in the field of animal fibers, e.g. hair and wool.

Speakman[1] has developed a non-destructive mechanical test with which he and others have investigated many of the changes resulting from treatment of wool fibers. The basis of the method is the fact that if a wool fiber is rapidly stretched to not more than 30 percent extension in cold water, contracted and allowed to rest 24 hours in cold water before re-extending, the second force-extension curve will closely duplicate the first. Thus, if a chemical reaction were carried out on the fiber just before the second test, the difference between the first and second force-extension curves would be due solely to the effect of the chemical reaction. The initial extension to 30 percent coupled with the rest in water was a non-destructive test. The advantages of such a test are very great; since both measurements are made on a single fiber, the need for cross-sectional area or denier measurements is eliminated and the problem of fiber-to-fiber variance is largely circumvented.

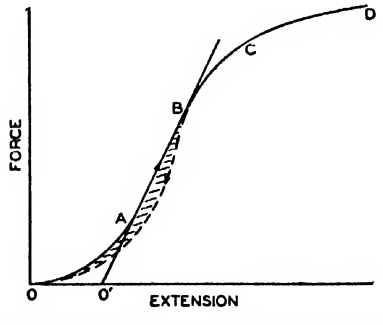


FIG. 1. Typical force-extension curve in the low range of extensions.

One objection to the test described above is the long time necessary between successive measurements. Furthermore, in order to follow the change of mechanical properties with time of treatment, a series of fibers must usually be used with the introduction of errors due to fiber-to-fiber differences. Speakman[1] points out that 30 percent extension is merely a general limit applicable to most animal fibers, and that for special conditions less drastic extensions might be needed. In the present work attention was focused on the Hooke's law region of the force-extension curves. Figure 1 shows a typical force-extension curve for a single wool fiber. In the region O–A the crimp of the fiber is being removed, in the region A–B (Hooke's law region) the force is linear with extension, in the region B–C the curve begins to depart from linearity and at some position, C, the yield point is reached. Extension of the fiber is usually measured from point O', and C generally occurs at from one to two percent extension. The slope of the curve in the linear region A–B will be called the Hookean slope. The fiber is not acting like a true Hooke's law spring in this region, since the contraction curve (dotted line) is always of initially steeper slope than the extension curve and a hysteresis loop (cross hatched) is formed. The Hookean slope is, however, a perfectly definite property that can be measured quite accurately. Dividing it by the cross-sectional area of the fiber will give a quantity analagous to Young's modulus.

Woods[2] has measured the Hookean slope for wool and hair as a function of relative humidity. He states that "under a very wide range of conditions an extension of the fibre by one percent or so involves no change in the fibre properties." The conditions under which the Hookean slope measurement is non-destructive were more closely investigated in the present work.

EXPERIMENTAL TECHNIQUE

Measurements of all mechanical properties were made with the Instron Tensile Tester.[3] Usually, the fiber was

* Research Fellow of the Textile Research Institute.

[1] J. B. Speakman, J. Textile Inst. 38, T102–126 (1947).

[2] H. J. Woods, Proc. Leeds Phil. and Lit. Soc. (Sci.) 3, 577–583 (1940).

[3] H. Hindman and G. S. Burr, Trans. Am. Soc. Mech. Eng. 71, 789–796 (1949).

extended slightly into the region B–C, and immediately contracted. For sample lengths of one to three inches, the rate of extension used was 0.10 in. per minute. Depending upon the fiber type and conditions of testing, it was possible to make a measurement of the Hookean slope every 3 to 15 min. without any dependence of the Hookean slope on the number of determinations. The shorter time intervals were for water soaked fibers; as the relative humidity of the air surrounding the fiber decreased, the longer time intervals were needed. If the time intervals between measurements were too short, successive values of the Hookean slope would show a slight increase. The fact that both of the above observations are due to imperfect elasticity of the fiber, even when it is extended only one to two percent, is shown in Fig. 2. The time interval between extensions was three to four minutes, the amount of semi-permanent set caused by the first extension was only 0.125 percent, yet it caused an increase in the Hookean slope of 7.3 percent (i.e., from 33.1 to 35.5 g/cm). When a sufficient time interval was allowed, successive measurements were reproducible to about ±1 percent. Furthermore, the test is non-destructive when performed on chemically treated fibers, whereas an extension to 30 percent would often cause a permanent change in fiber properties. Therefore, it is possible to use a *single* fiber to follow the change of Hookean slope as a chemical reaction with the fiber proceeds, and the change in Hookean slope will be due solely to the effect of the reaction. Tobolsky and Andrews[4] have performed similar experiments with rubber, calling the method "intermittent relaxation."

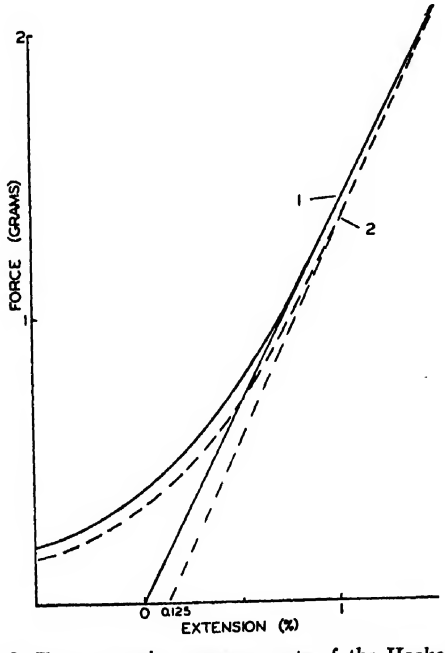


FIG. 2. Two successive measurements of the Hookean slope. (Three to four minute interval between extensions. R.H.=65 percent.) 1—First extension, Hookean slope=33.1 g/cm. 2—Second extension, Hookean slope=35.5 g/cm.

[4] A. V. Tobolsky and R. D. Andrews, J. Chem. Phys. 13, 3–27 (1945).

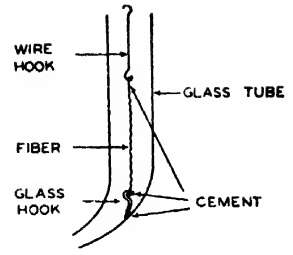


FIG. 3. Single fiber mounted for testing.

THE INTERACTION BETWEEN WOOL AND CONCENTRATED NEUTRAL SALT SOLUTIONS

As an example of the use of the non-destructive Hookean slope technique, some work on the heterogeneous reaction between single wool fibers and concentrated neutral salt solutions will be described.

Sookne and Harris[5] have shown that the mechanical properties of wool fibers immersed in concentrated solutions of neutral salts are not the same as the properties when in an atmosphere in equilibrium with these solutions. Neither are they the same as the properties of the fiber in pure water. In the present work, fibers at different initial conditions were treated with various concentrated neutral salt solutions. The change of Hookean slope with time of treatment was measured, and was plotted as k/k_0 *vs.* time, where k is the Hookean slope at any time after the reaction has started, and k_0 is the Hookean slope at zero time. The dimensionless quantity (k/k_0) is Woods' "relative Young's modulus."[2]

The fibers studied were from a sample of Columbia ewes wool, 56's grade, which had been cleaned by successive ether extractions and water washings. Single fibers were mounted in suitable containers by means of small hooks and de Khotinsky cement (Fig. 3). The container was attached to the Instron crosshead, the wire hook to the Instron load cell. All values of the Hookean slope were corrected for the small error introduced by the deflection of the load measuring element of the Instron Tester.

The curves of relative modulus (or relative Hookean slope) *vs.* time, for initially wet fibers that were treated with saturated salt solutions, are given in Fig. 4. The following explanation is proposed:

(a) Since the diffusion of water into, and out of, the fiber is probably much more rapid than the diffusion of salt ions, the first process to occur on immersing the fiber in a salt solution will be a dehydration or hydration of the fiber (depending upon its initial state) to a state in equilibrium with the activity of water in the salt solution. For the present purposes, the activity of the water can be indicated by the relative humidity over the solution. Woods[2] and Speakman[6] have shown that

[5] A. M. Sookne and M. Harris, American Dyestuff Reporter 26, 659–666 (1937); 27, 171–173 (1938).

[6] J. B. Speakman, J. Soc. Chem. Ind. 49, 209T (1930).

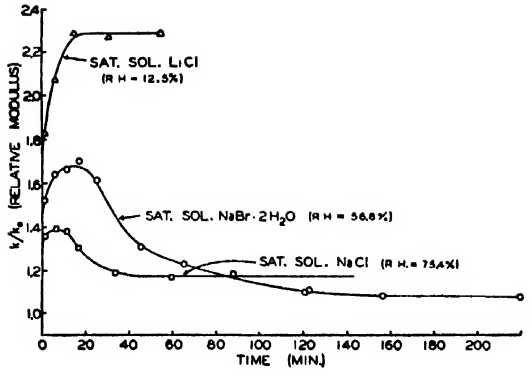


FIG. 4. Relative modulus *vs.* time after adding saturated salt solutions to an initially wet fiber.

dehydration of the fiber will cause the Hookean slope to increase.

(b) Concurrently, a slower process of absorption of the salt ions by the wool fiber occurs. These ions form lyospheres around charged centers in the fiber which are probably the acidic and basic side chains. Lundgren and co-workers[7] have proposed a very similar scheme to account for the effect of salts on synthetic feather keratin fibers. Absorption of the salt ions decreases the Hookean slope by either breaking salt linkages, or increasing the separation of the long polypeptide molecular chains of which the wool is composed. Both mechanisms may be important.

(c) If the activity of water in the salt solution is very low, the consequent rapid dehydration of the fiber may reduce the swelling to such an extent that the slower subsequent diffusion of the salt ions into the wool fiber is greatly hindered. Speakman[8,9] has shown that a wet fiber will absorb molecules that are too large to diffuse into a dry fiber.

Thus, when the fiber was treated with a saturated solution of either sodium chloride or sodium bromide (Fig. 4) the first process was one of dehydration causing the relative modulus to increase. Note that, as expected, the lower the relative humidity over the solution, and thus the greater the dehydration, the greater was the initial increase in relative modulus. The concurrent

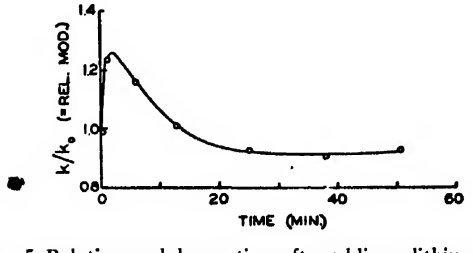


FIG. 5. Relative modulus *vs.* time after adding a lithium chloride solution (one vol. saturated solution, 2 vol. water) to an initially wet fiber.

[7] Lundgren, Stein, Kearn, and O'Connell, J. Phys. and Colloid Chem. **52**, 180–206 (1948).
[8] J. B. Speakman, J. Soc. Chem. Ind. **48**, T321–324 (1929).
[9] J. B. Speakman, Proc. Roy. Soc. **132A**, 167–191 (1931).

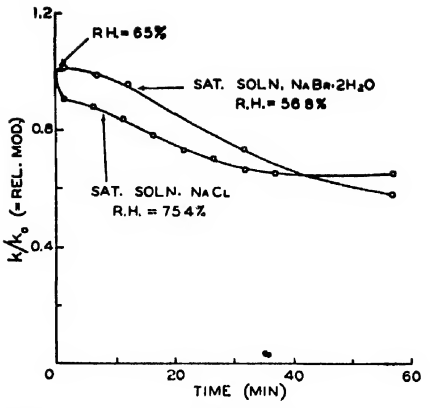


FIG. 6. Relative modulus *vs.* time after adding saturated salt solutions to a fiber initially at 65 percent R.H.

slower absorption of salt ions caused the decrease after the maximum. The bromide causes a greater decrease than the chloride, but the reason for this is not clear. The upper curve, obtained with the saturated lithium chloride solution with a very low partial pressure of water vapor over it, shows no evidence of the absorption of salt although some absorption undoubtedly took place in the early stages of dehydration before the molecular network was sufficiently dehydrated to forbid further diffusion of the salt ions. It might be pointed out that all the curves of Fig. 4 were obtained on the same fiber. It had previously been found that prolonged washing in water would remove all of the absorbed salt, and leave the fiber in its initial state. As might be expected, when a fiber in equilibrium with a saturated sodium bromide or sodium chloride solution was immersed in water, the relative modulus-time curve was found to go through a minimum as first water diffused into the partially dehydrated fiber and then the salt ions diffused out.

For an unsaturated lithium chloride solution, the dehydration of the fiber should not be great enough to prevent absorption of the salt ions, and a maximum in the relative modulus-time curve should be observed. In Fig. 5 this effect is shown for a different fiber from the one illustrated in Fig. 4. Note that, for this concentration of lithium chloride the Hookean slope finally decreased below the initial value in water. This action might be expected, since Alexander[10] has pointed out that lithium salts will cause hydrogen bond breakdown, which would be in addition to the salt effect described above.

When a fiber initially conditioned at 70°F, 65 percent R.H. was brought into contact with a saturated sodium chloride solution the relative modulus-time curve (Fig. 6) decreased rapidly at first as water was absorbed to bring the fiber to equilibrium with the higher activity of water, corresponding to a relative humidity of 75.4 percent, in the salt solution. This decrease was followed

[10] P. Alexander, Research **2**, 246–247 (1949).

by another slower one as salt ions were absorbed. When, however, a different fiber initially at 65 percent R.H. was immersed in a saturated sodium bromide solution (R.H. = 56.8 percent) the relative modulus-time curve went through a slight maximum as the two effects of slight further dehydration of the fiber and absorption of salt ions opposed each other. Eventually, the relative modulus decreased as the ions were absorbed. When this fiber was then removed from the bromide solution and allowed to come to equilibrium with a 65 percent R.H. atmosphere, a slight further decrease of the relative modulus was noted. The latter decrease resulted from the fact that water could be absorbed from the atmosphere (which was at a higher relative humidity than that corresponding to the saturated sodium bromide solution), but the salt ions could not escape. On washing the fiber with water and once more allowing it to stand at 65 percent R.H., the initial value of the Hookean slope was recovered, i.e. the relative modulus became 1. The experiment just described proves that the fiber absorbs salt ions from the solution and that this absorption affects its mechanical properties.

In Fig. 7 are given the relative modulus *vs.* time curves for a series of initially wet fibers which had been treated with a saturated sodium bromide solution. The striking fact is the large fiber-to-fiber variance. It is obvious that, lacking the Hookean slope non-destructive test wherein a *single fiber can be used to follow the entire course of the reaction*, a relatively large number of samples would have had to be tested to clearly indicate the reaction mechanism.

THE INTERACTION BETWEEN WOOL AND GLYCERINE SOLUTIONS

The behavior of the fiber in concentrated salt solutions is in many respects similar to what one would expect in aqueous solutions of large polar organic molecules. Speakman[9] has described such systems (e.g. glycerine-water-wool). If a dry fiber is treated with the anhydrous liquid, the molecular network of wool is too tight for many glycerine molecules to diffuse into the fiber. If, however, an aqueous solution of glycerine is used, the water swells the fiber to an extent permitting absorption of the glycerine molecules. In the first case,

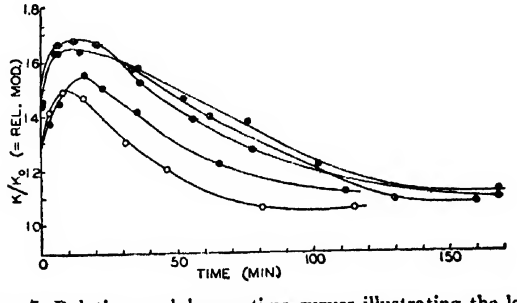


FIG. 7. Relative modulus *vs.* time curves illustrating the large fiber-to-fiber variance (initially wet fibers treated with a saturated sodium bromide solution).

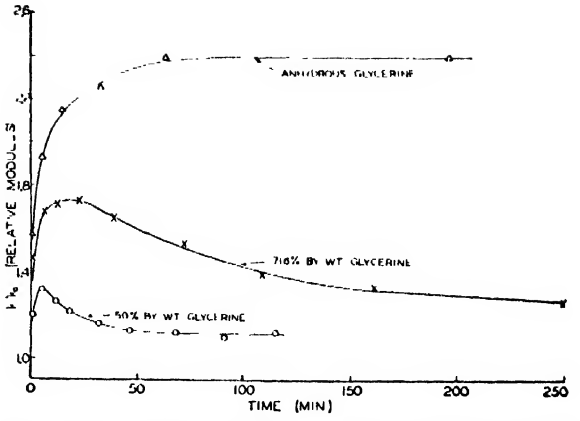


FIG. 8. Relative modulus *vs.* time after adding glycerine solutions to an initially wet fiber.

the behavior should correspond to that for a saturated lithium chloride solution. In the second case, on adding an aqueous glycerine solution to an originally wet fiber, the relative modulus-time curve should go through a maximum. The data of Fig. 8 bear out these hypotheses. The curves for the aqueous solutions are for the same fiber, but that for the anhydrous glycerine is for a different fiber.

USE OF OTHER MECHANO-CHEMICAL METHODS

Extension to 30 Percent

The non-destructive test involving a 30 percent extension was also used for a comparison with the results using the Hookean slope. Figure 9 shows force-extension curves on a human hair fiber extended 30 percent of its original length. A 24-hour rest period in water was allowed to elapse between each extension. Curve *a* is the initial extension in water. Curve *b* is an extension after the fiber had been standing in a saturated sodium chloride solution for 42 min. Curve *c* is a re-extension in water after the salt had been washed out. The reaction occurring in *b* was reversible, and the fiber could in this case be reused. Curve *d* resulted after the fiber had been standing in saturated salt solution for 15.5 min. It is evident that the curve of energy to extend 30 percent *vs.* time must go through a maximum. Also, if the amount of semi-permanent set resulting from extending to 30 percent and then contracting at the same rate were plotted against time, the curve would go through a maximum. Both of these facts are in accord with the explanation developed by using the Hookean slope measurement.

Repeated Cycling Technique

One further mechanical testing technique will be mentioned. When a wet wool fiber is subjected to repeated extensions to a fixed elongation, the force-extension curve is shifted progressively lower on the force axis, (Fig. 10). (The rate of extension used here was about 50

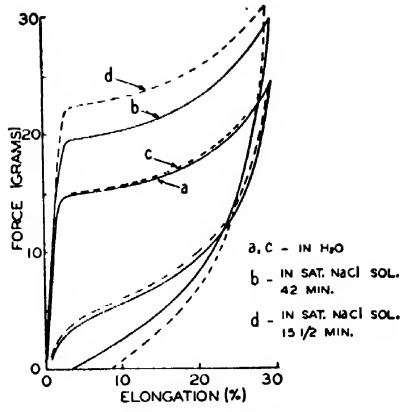


FIG. 9. Force-extension curves to 30 percent extension.

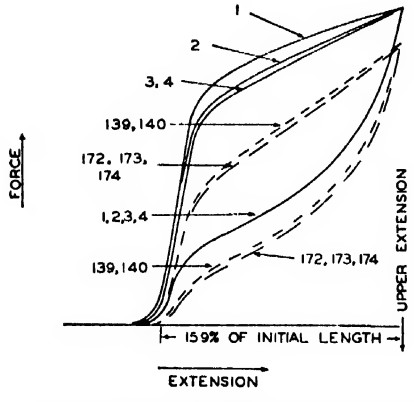


FIG. 10. Successive extension-contraction cycles on a wet wool fiber.

percent per minute.) After a number of cycles, however, the rate of change with number of cycles becomes very slow (compare cycles 140 and 172 of Fig. 10). If, at such a point, some substance which interacts with wool is introduced, the entire change in the force-extension curve should be due substantially to the effect of the extraneous substance on the wool fiber. With the Instron Tester the repeated cycling of the fiber can be made automatic. There are, however, several limitations to this method. (a) The effect of the reaction on a mechanically degraded, rather than the initial undamaged fiber, is being studied. (b) The continued cycling may cause further breakdown due to the repeated stressing of the reacted fiber, whereas it would not do so with the original fiber. This may obscure the main value of the experiment which is to measure the progressive change of mechanical properties of wool with time of reaction. (c) The force-extension curves become somewhat difficult to interpret if practically 100 percent elasticity is not displayed.

In spite of these limitations the method has some value as shown by the results of its application to the interaction between wool and a saturated salt solution (Fig. 11). The repeated cycling technique gave much the same information as the non-destructive extension to 30 percent method in a fraction of the experimental time. The saturated salt solution was sodium chloride and the fiber was subjected to 175 extension-contraction cycles before the solution was added.

Swelling Experiments

In accord with the results of the mechano-chemical experiments, White and Barnard[11] of this laboratory have shown that, when an initially wet hair fiber is immersed in a concentrated salt solution, the swelling *vs.* time curve goes through a minimum. When very concentrated solutions such as saturated lithium bromide are used, the swelling decreases and does not thereafter increase.

WORK IN PROGRESS

If a fiber is initially conditioned to the relative humidity existing over a salt solution before being

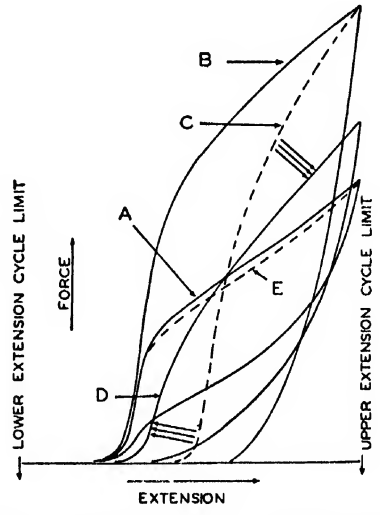


FIG. 11. Repeated cycling technique applied to the interaction between a wool fiber and a saturated sodium chloride solution. A—in water before adding salt solution; B—first cycle after adding sodium chloride solution; C—second cycle in salt solution; D—54 min. after adding salt solution; and E—first cycle after replacing salt solution with water.

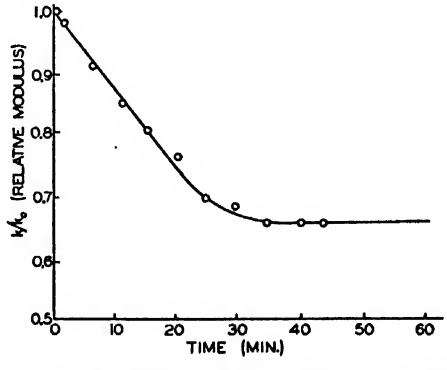


FIG. 12. Relative modulus *vs.* time after adding a saturated sodium chloride solution to a fiber initially at 75.4 percent R.H.

[11] H. J. White, Jr. and W. S. Barnard (unpublished communication).

treated with the solution, the initial hydration or dehydration will be eliminated, and the only effect measured should be that of the absorption of the salt ions. Figure 12 shows the relative modulus-time curve for a fiber, initially conditioned in a stream of air which had been bubbled through a saturated sodium chloride solution, and then immersed in a saturated sodium chloride solution. By making use of such curves, the rate of absorption of sodium chloride from its saturated solution is being studied as a function of animal fiber type and diameter.

SUMMARY

(1) An investigation has been made of the conditions under which the slope of the force-extension curve in the Hooke's law region may, for animal fibers, be non-destructively measured.

(2) A *single* fiber can be used to investigate a chemical reaction by measuring the change of Hookean slope with time of reaction.

(3) The usefulness of this technique is illustrated by its application to the heterogeneous reaction between wool and concentrated neutral salt solutions. The heterogeneous reaction between wool and aqueous glycerine solutions was also studied.

(4) Other mechano-chemical methods as well as swelling experiments are described. They give results consistent with those for the Hookean slope technique.

(5) A mechanism consistent with the experimental results is presented.

ACKNOWLEDGMENT

The author wishes to thank Professor J. C. Whitwell under whose direction this work was done and Dr. J. H. Wakelin for their many helpful suggestions. It is also a pleasure to acknowledge the support, in the later stages of this work, of the various organizations sponsoring the "International Wool Project" being carried out in the Textile Research Institute Laboratories. These organizations include The Wool Bureau, Inc., representing both the American Wool Council and the International Wool Secretariat, the U. S. Department of Agriculture, the Office of Naval Research and a group of American wool manufacturers.

Equivalent Effects of Time and Temperature in the Shear Creep and Recovery of Elastomers

F. S. Conant, G. L. Hall, and W. James Lyons

Chemical and Physical Research Laboratories, Firestone Tire and Rubber Company, Akron Ohio

An explicit relationship is set forth for the time-temperature dependence of the *viscoelastic* phenomena in the superelastic polymers. An empirical equation that was found to represent adequately the above-mentioned relationship over the entire multiple-temperature curve is of the form: $\log t_e = [C_\alpha/(T-b)]+C_\beta$. Experimental verification is given for the equivalent influence of time and temperature on the creep and recovery of compounds based on Hevea, GR-S, Neoprene GN, Butaprene, and butyl rubber. A comparison of the empirical equation with that of a theoretical reaction-rate equation of Tobolsky and Eyring indicates a temperature dependence of the energy of activation.

INTRODUCTION

THE dependence of creep phenomena on temperature in high polymers is well known, having been observed and discussed by a number of investigators.[1] The experiments show that, for a given time interval, the viscoelastic strain (or recovery) tends generally to rise to a higher level with an increase in the temperature of the experiment. The present series of experiments have disclosed that for rubber-like materials a consistent elevation of the strain-time curves with an increase in temperature is obtained, if corrections are made for the Gough-Joule effect. While a number of empirical relationships describing the dependence of *viscous* flow on temperature have been proposed and analyzed theoretically, no clear-cut, explicit relationships have been set forth for the temperature-dependence of *viscoelastic* phenomena in the superelastic polymers. The closest approach to such a relationship appears to be one based on a temperature function for viscosity, quoted by Leaderman.[1]

The observed nature of the dependence of viscoelastic behavior on temperature implies that it should be experimentally possible to duplicate the strain S_v attained in time t_1 at temperature T_1 with an equal strain S_v attained in time t_2 at T_2. As has been indicated, a series of creep or recovery experiments on a given material, each conducted over the same time interval, but at different T, will yield a family of creep (or recovery) curves; for increasing values of T the curves will lie higher and higher along the S direction. From these considerations emerges the possibility of selecting a series of temperatures for these experiments such that each curve at its termini in the selected time interval will match the termini of the curves obtained at the next higher and lower temperatures. Each terminus provides

[1] For a review of these studies see H. Leaderman, *Elastic and Creep Properties of Filamentous Materials and Other High Polymers* (Textile Foundation, Washington, 1943), pp. 30–33, 75 ff.

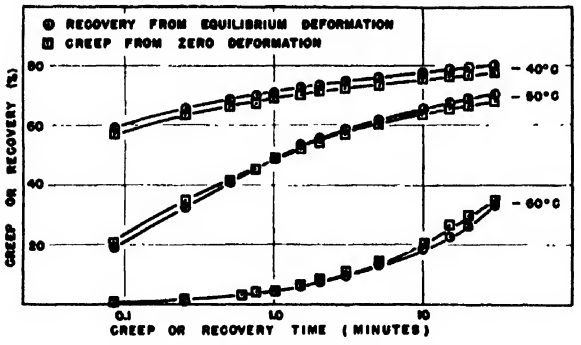


FIG. 1. Equivalence of creep and recovery, Hevea, 35.6 lb /sq. in. load.

the *particular* S_e which, given by one combination t_1 and T_1, can be duplicated by another combination t_2 and T_2. Plotted end-to-end the curves will become segments of one over-all multiple-temperature curve. Tobolsky and Andrews[2] observed such a continuous relationship in analogous stress relaxation tests and proposed an implicit relationship among time, temperature, and stress. Without a knowledge of the time-temperature function, however, it does not appear possible to predict whether the general, synthetic graph will be a smooth curve without discontinuities at the junctions of the segments.

With the aim of testing the foregoing theory, as well as deriving an analytical expression for the relationship between time and temperature, observations in creep and recovery on Hevea, GR-S, Neoprene GN, Butaprene, and Butyl rubber were undertaken.

APPARATUS AND METHODS

The test apparatus used has been described previously.[3] Essentially, it consists of a means of applying a shear stress to a double sandwich-type of specimen, and of observing the deformation of the specimen as test conditions are varied. The test specimens used were also based on the same compounding formulas and cures as given in the above reference.

In the creep tests performed, the unstressed test specimen was first brought to thermal equilibrium at the desired temperature. Twenty minutes conditioning time was allowed after a thermocouple in the center metal web of the specimen indicated that the test temperature had been reached. A shear stress of 35.6 p.s.i. was then applied quickly and the consequent deformation noted at appropriate intervals from 5 sec. to 30 min. after loading. In general, the measurements were made first at the lowest test temperature. After completion of the test the load was removed and the temperature of the test chamber raised. Recovery was usually found to be complete, after conditioning at the next higher test temperature.

In the recovery test the specimen was conditioned overnight (16 hr.) at 35°C under load. At the end of this time, deformation had proceeded to essentially an equilibrium state. The temperature was then gradually lowered to the lowest test temperature, during which time the deformation increased because of the Gough-Joule effect. A curve of equilibrium deformation *vs.* temperature was obtained during the cooling. As soon as thermal equilibrium was reached at the lowest test temperature the load was removed and the recovery noted at appropriate intervals from 5 sec. to 30 min. At the end of this time the load was reapplied and the temperature raised to the next test temperature. The reestablishment of equilibrium deformation was shown by the approach of the deformation to that of the equilibrium curve previously obtained.

In a study of creep or recovery it is often advantageous, if not necessary, to compensate for the Gough-Joule effect, through a normalization of the data. Unless this is done it is found that after a sufficient creep or recovery time the deflection at a low temperature exceeds that at a higher temperature, because of

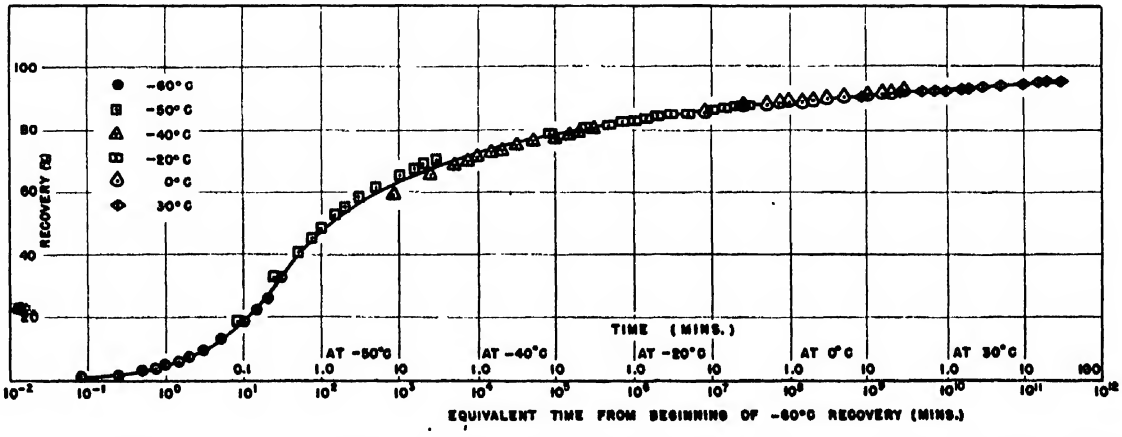


FIG. 2. Recovery from equilibrium deformation, Hevea, 35.6-lb./sq. in. load. Test temperatures taken at multiples of 10°C.

[2] A. V. Tobolsky and R. D. Andrews, J. Chem. Phys. **13**, 3 (1945).
[3] Conant, Hall, and Thurman, J. App. Phys. **20**, 526 (1949).

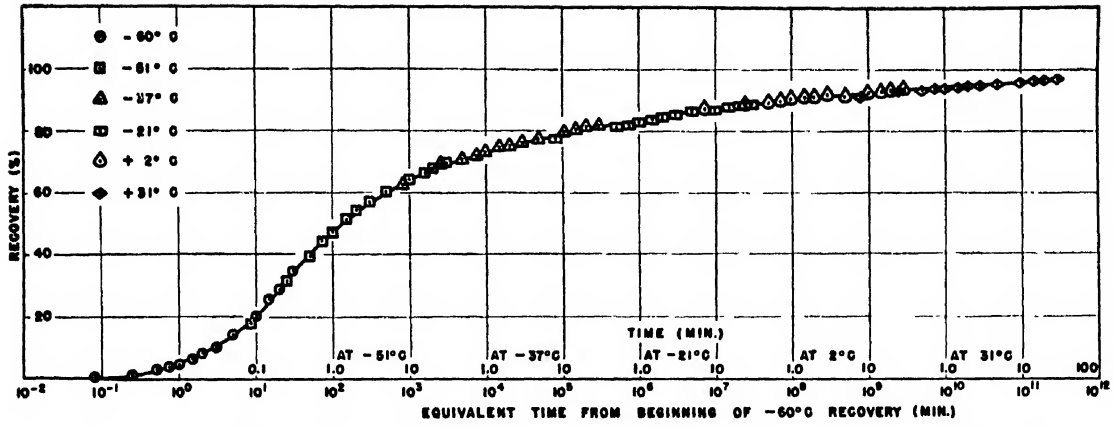


FIG. 3. Recovery from equilibrium deformation, Hevea, 35.6-lb./sq. in. load; adjusted temperature intervals.

this effect. The data presented here were normalized by presenting each creep datum as a percentage of the equilibrium deformation for that temperature and each recovery as a percentage of complete recovery.

Both creep and recovery curves were obtained on all of the stocks in the present series. The normalized data for each showed the two to be equivalent within the experimental error of either Figure 1 shows a typical example of the close correspondence of the two types of data. Only the recovery data will be presented for the other stocks in the series.

JOINING OF RECOVERY CURVES OBTAINED AT DIFFERENT TEMPERATURES

From data such as those shown in Fig. 1, it appears that the creep or recovery curves at different temperatures might be joined end-to-end to form such a continuous, synthetic curve as was projected in the introduction. Such a continuous curve is meaningless, of course, unless the segments are matched so that the logarithmic time scale is continuous. Brief reflection will disclose that it is not necessary actually to procure a continuous locus of experimental points (such as could be obtained only by a fortuitous choice of temperatures) in order to secure a smooth multiple-temperature curve. When a series of parallel curve segments, obtained at reasonably well-scattered temperatures, and overlapping on the time scale, are available, the general continuous curve can be derived by drawing it parallel to these segments. This procedure is illustrated in Fig. 2 which shows 30-min. recovery data for Hevea stock B. The data at some of the temperatures are seen to be displaced from the smooth continuous curve.

The smooth curve of Fig. 2 strongly implies the existence of a continuous functional relationship between normalized creep or recovery time and temperature. Tentatively assuming its correctness, the curve may be used to supply numerical data for a set of simultaneous equations to test various t-functions of T. Following this procedure an empirical equation of the following form has been found to represent adequately the data over the entire viscoelastic region.

$$\log t_c = [C_\alpha/(T-b)] + C_\beta, \qquad (1)$$

where t_c = creep or recovery time, T = absolute temperature, and C_α, b, and C_β are arbitrary constants. The resemblance of Eq. (1) to the Arrhenius formula for the temperature-dependence of rate processes, will be recognized.

While an estimate may be made from data such as those in Fig. 2, of the temperatures at which tests should be made in order that the experimental curve segments should be on the continuous line, it is preferable, by way of testing Eq. (1), to use it to derive the appropriate temperatures. Using the equation and the data in Fig. 2, temperatures of −60°, −51°, −37°, −21°, +2°, and 31°C were selected for test. Figure 3 shows that recovery curves obtained at these selected temperatures do, indeed, form a continuous multiple-temperature curve.

An additional confirmation of the equivalence of time and temperature is provided by the data in Fig. 4. The experimental points show the recovery at −60°C, while the curve shows the joined recovery curves of the same specimen (Hevea stock B). The two are seen to be very nearly equivalent, for the 16-hr. period of test at −60°C.

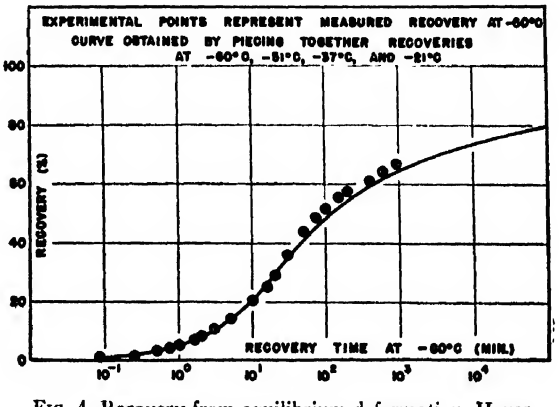


FIG. 4. Recovery from equilibrium deformation, Hevea, 35.6-lb./sq. in. load.

TABLE I. Values of empirical parameters in Eq. (1) for 50 percent recovery from equilibrium shear deformation of various elastomers.

Stock	C_α	C_β	b	Derived test temp. (°C)
Hevea	1780	−17.5	122	−60, −51, −37, −21, 2, 31
GR-S	3121	−19.1	100	−50, −40, −27, −20, −4, 30
Neoprene GN	480	−7.52	193	−40, −32, −20, 0, 40[a]
Butyl rubber	5040	−23.52	33	−60, −46, −38, −30, 0, 40
Butaprene	720	−8.82	183	−30, −18, −10, 0, 30

[a] This is the derived test temperature for the material in the same crystalline state as it was at lower temperatures. In the test made at 30°C some of the crystals had probably melted, resulting in a higher recovery than that calculated. This same singularity appeared in the creep curves for Neoprene GN.

A continuous multiple temperature recovery curve for a GR-S stock is shown in Fig. 5. Test temperatures were selected by an application of Eq. (1) to previous data obtained at even decade temperature intervals. This curve is quite different in shape from that in Fig. 3 for Hevea, reflecting the more sluggish response of GR-S to low temperature stress.

Figures 6–8 show joined together recovery curves at even decade, temperature intervals on motor mounting-type stocks of Neoprene GN, Butyl rubber, and Butaprene. (Formulas are given in reference 3.) Each curve is characteristically sigmoidal. Only the Neoprene GN, however, shows the sharp increase in deformation with time that was evident in the Hevea stock. The Neoprene GN curve also shows a slight upturn at high temperatures, possibly because of the disappearance of a slight degree of crystallization.

An application of the above empirical formula to the data is shown in Table I which gives the constants and derived test temperatures. These calculations were based on the time required for 50 percent recovery at each temperature. The constant C_β is affected by the degree of recovery chosen. The constants C_α and b appear to be properties of the material tested and are independent of the reference point. Values of T and $\log t_e$ were substituted in Eq. (1) at three different temperatures for which the data were on the smooth curve. Solution of the resulting simultaneous equations gave the constants C_α, C_β, and b. The derived test temperatures given in Table I were obtained by an application of Eq. (1) to data obtained at even decade, temperature intervals. If the data shown in Figs. 6–8 had been obtained at the derived test temperatures, then all experimental points should have been on the curves, as they were in Figs. 3 and 5.

The interdependence of time, temperature, stress, and strain in the viscoelastic phenomena is well known. Mullins[4] recognized the need for a test that would use time as a parameter in determining low temperature serviceability. He also stated the need for a simple expression relating time with modulus. Many of the low temperature tests, such as Gehman's torsion test,[5] the Young's modulus test of Conant and Liska[6] and the cold compression set test[7] recognize time only to the extent that they specify a particular loading time in making their measurement. The above-mentioned tests are valuable in that they give the temperature below which, because of change in a particular physical property, the product can no longer be considered serviceable. However, time is not directly included as a parameter in these tests. The T-50 test[8] is arbitrary in that both time and temperature are varied simultaneously. Equation (1) does give a simple relationship between time and temperature, thus enabling time to be used as a parameter in determining low temperature serviceability.

The applicability of the data in Table I in conjunction with Eq. (1) to determine a low temperature serviceability index is readily apparent. For example, one can obtain the temperature at which a stock will recover 50 percent in 1 min. On this basis the stocks tested give the

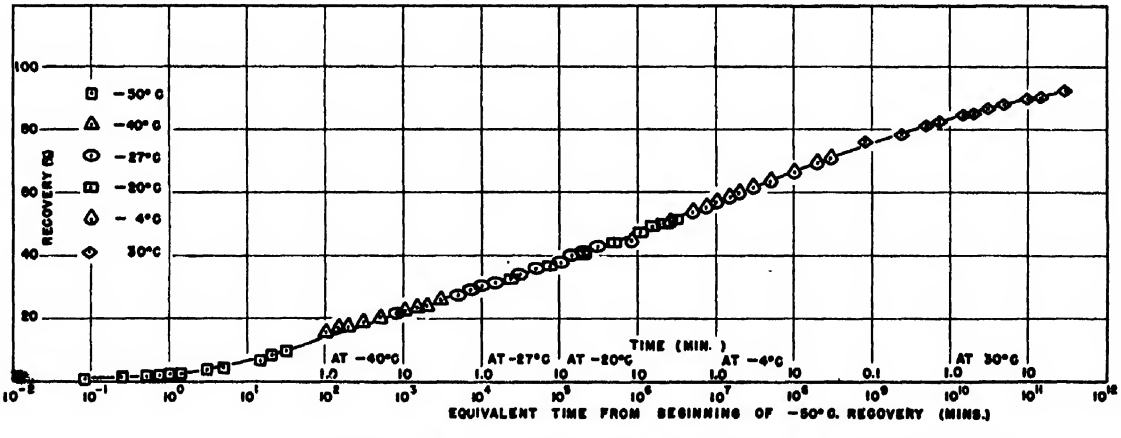


FIG. 5. Recovery from equilibrium deformation, GR-S, 35.6-lb./sq. in. load; adjusted temperature intervals.

[4] L. Mullins, Trans. I. R. I. 21, 247 (December, 1945).
[5] Gehman, Woodford, and Wilkinson, Ind. Eng. Chem. 39, 1108 (1947).
[6] F. S. Conant and J. W. Liska, J. App. Phys. 15, 767 (1944).
[7] Morris, Hollister, and Mallard, India Rubber World 112, 455 (July, 1945).
[8] W. S. McCortney and J. V. Hendrich, Ind. Eng. Chem. 33, 579 (1941).

following indices: Hevea −49.5°C; GR-S −10°C; Neoprene GN −16°C; Butyl rubber −26°C; and Butaprene −6°C. The particular conditions selected for defining the index would, of course, be determined by the use for which the rubber was intended.

THEORETICAL CONSIDERATIONS

In an application of the reaction-rate theory of the mechanical properties of polymers to creep under constant stress, Tobolsky and Eyring[9] arrive at the equation

$$S=\frac{f}{G_1}+\frac{1}{B_2G_1}\ln A_2B_2G_\mu+\frac{1}{B_2G_1}\ln t, \qquad (2)$$

where S= strain in the specimen; f= total stress; G_1 and G_μ are appropriate elastic moduli associated with strains of the primary and secondary network bonds; and t= duration of creep experiment. The quantities A_2 and B_2 are defined by the equations

$$A_2=2n_2\lambda_2(kT/h)\exp(-\Delta F/kT), \qquad (3)$$

and

$$B_2=\lambda_2/2N_2kT, \qquad (4)$$

where k and h are Boltzmann's and Planck's constants, respectively; T is the absolute temperature; F is the free energy of activation for the deformation process; and n_2, λ_2, and N_2 are theoretical quantities related to the unit or elementary process.

If Eq. (2) is written out in full and certain terms are canceled, where possible, we arrive at the equation

$$S=\frac{f}{G_1}+\frac{2NkT}{G_1\lambda_2}\left[\ln\left(\frac{2n_2\lambda_2}{h}\cdot\lambda_2\frac{G_\mu t}{2N_2}\right)-\frac{\Delta F}{kT}\right]. \qquad (5)$$

k and h are well recognized as universal constants. Furthermore, we shall accept the material and structural characteristics, G_1, G_μ, λ_2, and ΔF as being time- and temperature-independent. The parameter N_2, however, Tobolsky and Eyring have deduced, decreases with a rise in temperature, though they do not give an expression for the relationship. We shall assume that the inverse variation of N_2 with absolute temperature is a

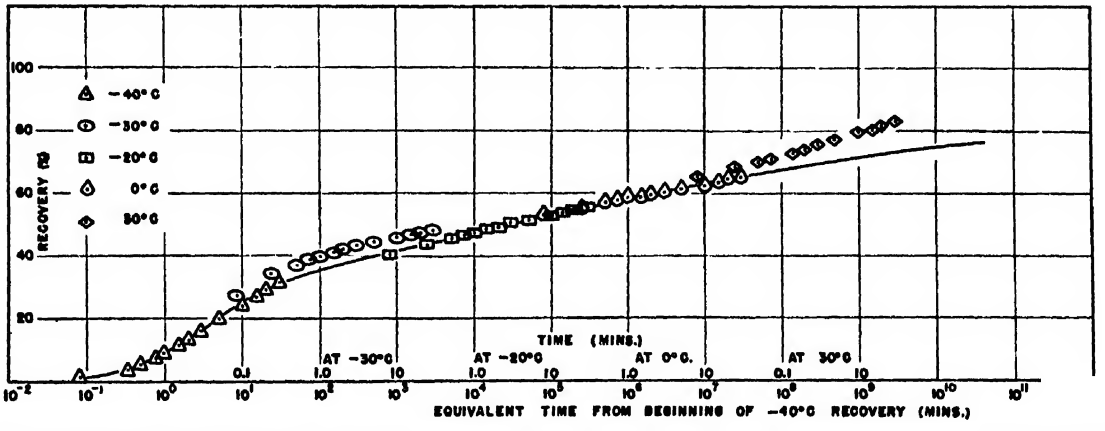


FIG. 6. Recovery from equilibrium deformation, Neoprene GN, 35.6-lb./sq. in. load. Test temperatures taken at multiples of 10°C.

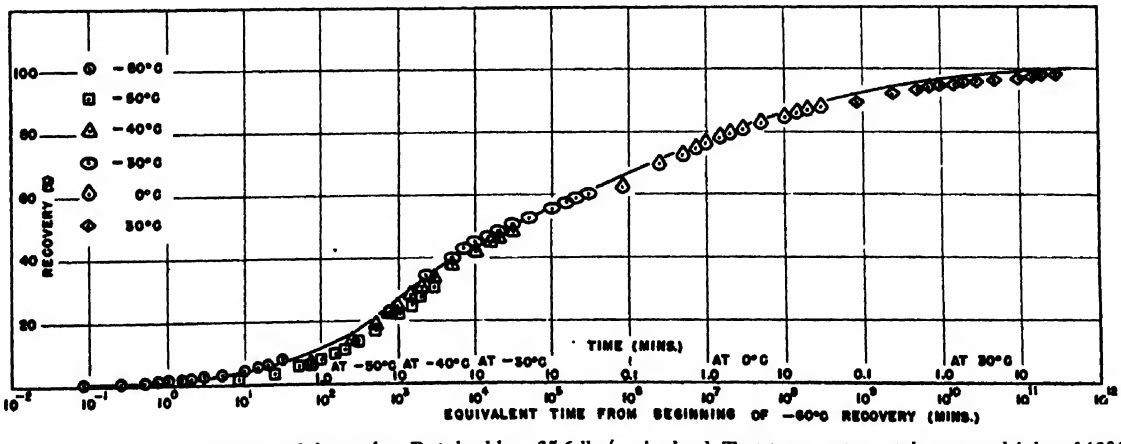


FIG. 7. Recovery from equilibrium deformation, Butyl rubber, 35.6-lb./sq. in. load. Test temperatures taken at multiples of 10°C.

[9] A. Tobolsky and H. Eyring, J. Chem. Phys. 11, 125 (1943).

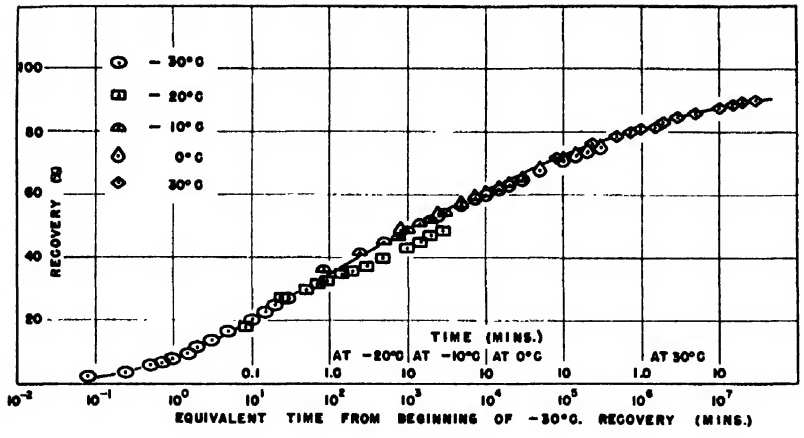


FIG. 8. Recovery from equilibrium deformation, Butaprene, 35.6-lb./sq. in. load. Test temperatures taken at multiples of 10°C.

good approximation, and also, that n_2 effectively varies inversely as the temperature over the range covered by the present experiments. The assignment of this functional dependence to N_2 and n_2 is admittedly quite arbitrary, from the theoretical point of view. However, if the network of secondary bonds can be regarded as behaving like an ideal gas, the assigned dependence on T becomes entirely reasonable, for N_2 represents the number of secondary bonds per unit area, while n_2 is the number of elementary deformation processes per unit length.*

Thus we obtain for the viscoelastic component of the deformation

$$S_v \equiv S - f/G_1 = C_1 \log C_2 t - C_3/T \tag{6}$$

involving the three parameters

$$\begin{aligned} C_1 &= \text{const.} \cdot 2.30(2k/G_1\lambda_2), \\ C_2 &= \text{const.} \cdot \lambda_2^2 G_\mu/h \end{aligned} \tag{7}$$

and

$$C_3 = \text{const.} \cdot 2\Delta F/G_1\lambda_2.$$

By rearranging terms, and selecting a fixed value for S_v, Eq. (6), placing $t = t_c$, may be written with a new set of constants

$$\log t_c = (C_\alpha/T) + C_\beta \tag{8}$$

where

$$C_\alpha = C_3/C_1; \quad C_\beta = (S_v/C_1) - \log C_2.$$

Equation (8) is seen to bear a close resemblance to the empirical Eq. (1), the difference being in the absence of the "correction" term, b on the temperature T, which appears in Eq. (1). The presence of the b in the empirical equation may be interpreted as an indication that the activation energies for the deformation processes involved in these experiments are not independent of temperature, but rather, are functions of temperature, of a form which gives decreasing values with increasing temperature:

$$\Delta F = \text{const. } T/(T-b). \tag{9}$$

Evidence for the temperature-dependence of the activation energy for *flow* in liquids has been recognized by Ewell,[10] though in this case indications were that the activation energy was an increasing function of temperature.

ACKNOWLEDGMENTS

The constructive interest of Dr. George R. Thurman, and his discussions with the authors, have produced valuable suggestions. The encouragement of the officials of the Firestone research organization, and their cooperation in making possible this contribution, is gratefully acknowledged.

* It may be remarked that attempts, before Eq. (1) was adopted for trial, to fit to the experimental data a working equation based on the assumption of the temperature-independence of N_2 and n_2, proved unsatisfactory.

[10] R. H. Ewell, J. App. Phys. **9**, 252 (1938).

Studies in Plasticization of Polyvinyl Chloride
I. Physical Changes and Their Measurement

M. L. Dannis
B. F. Goodrich Research Center, Brecksville, Ohio

The softening of a hard resin, polyvinyl chloride, by a liquid, dioctyl phthalate, is a set of processes occurring in sequence. Imbibition of the liquid is accompanied by volume reduction in the system. Dilatometer methods evaluate the amount and rate of volume decrease. The rate is temperature dependent, activation energy about 80 kcal. Solution of liquid into the resin follows with no further volume, but large dielectric change. Dielectric changes are measured on material in a cylindrical condenser, using a Schering bridge, and are rate dependent, activation energy about 110 kcal. Data are best described, assuming that segments of the polymer molecule "react" independently of one another.

I. INTRODUCTION

PLASTICIZED polyvinyl chloride (Koroseal) is a material used widely in industrial or consumer goods because of its many excellent physical and chemical properties. Some of these properties such as electrical characteristics have been studied in detail,[1] since they were important to both the manufacturer and the consumer. In all studies published thus far, the material actually investigated was the final product, usually milled and pressed into the required physical form.

This paper concerns itself with a study of the reactions and processes that occur when this final material is made from the original raw materials. These changes are largely physical in nature, but do follow many of the rules of "rate-process" chemistry, permitting some evaluation of thermodynamic barriers in the reacting systems.

"Plasticization" is a term used to describe the softening of a hard resin, usually by means of a viscous liquid. Specifically, polyvinyl chloride, henceforth abbreviated PVC, which is a hard, brittle, horny material, may be softened to a resilient, rubbery material by the addition of appropriate amounts of esters such as di-2-ethylhexyl phthalate (DOP)[2] or *m*-tricresyl phosphate (TCP). The "reactions" accompanying this softening action are not simple and thoroughly understood, but at least three basic steps have been identified.

These steps are listed. (1) "Shrinkage"—believed to be a measure of plasticizer imbibition by the polymer. (2) "Dielectric Conversion"—believed to be a solution of the plasticizer into the resin. (3) "Fusion"—believed to be a simple physical consolidation.

II. SHRINKAGE

The specific volume of plasticized PVC is less than the volume of the components present.[3,4] This difference in volume has been studied as a rate-process phenomenon, using a dilatometer technique, described below.

About three grams of PVC powder or granules, as sold commercially, are weighed into a filter paper capsule. The capsule is inserted into a dilatometer bulb, followed by a glass plug. The plug partially protects the sample from overheating in subsequent glass sealing. The sample bulb is partially constricted just below the plug. Plasticizer fluid is put in the reservoir at the other end of the dilatometer. An oven is slipped over the bulb, vacuum connections made to both ends, and pumping begun. The oven is set at 60°C for one hour, and the liquid is gently boiled for a short time. The bulb is sealed off while the system is still evacuated, then cooled in ice water. The dilatometer is tipped, allowing the cool fluid to immerse the sample, still in ice water. After a few minutes the remaining

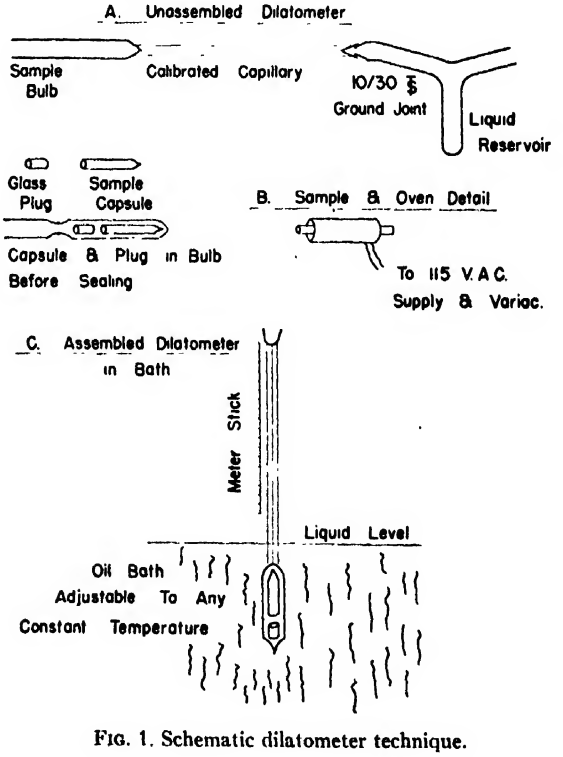


FIG. 1. Schematic dilatometer technique.

[1] Davies, Miller, and Busse, "Dielectric properties of plasticized polyvinyl chloride," J. Am. Chem. Soc. 63, 361 (1941).
[2] T. L. Gresham, U. S. Patent 2,325,951 (August 3, 1943).
[3] Raymond M. Fuoss, "Electrical properties of solids. V," J. Am. Chem. Soc. 61, 2334 (1939). The statement above Fig. 1 was later changed, to admit that slight volume loss in plasticizing PVC was always found.
[4] Raymond M. Fuoss, J. Am. Chem. Soc. 63, 2401 (1941).

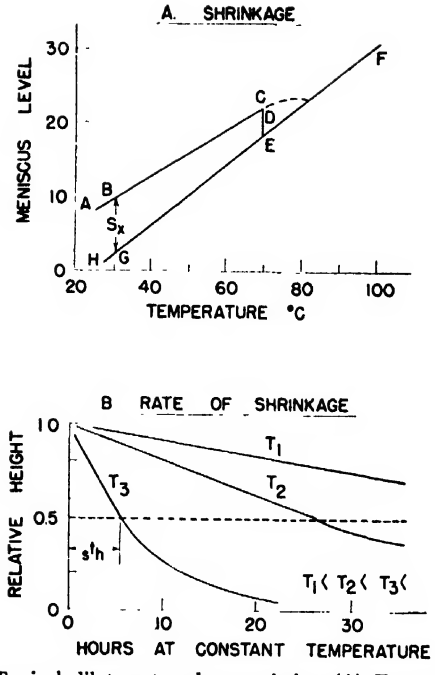


FIG. 2. Typical dilatometer characteristics. (A) Expansion and contraction. (B) Loss of volume at constant temperature.

vacuum connection is broken and the bulb removed from the ice bath. Excess fluid is sucked from the calibrated capillary by a smaller probe capillary. A meter stick is wired to the capillary, and the sample bulb put in a "constant" temperature bath. Typical apparatus is shown in Fig. 1, A, B, C.

The bath is held at a selected temperature long enough to establish thermal equilibration, about 5 minutes, and the meniscus level read. The bath temperature is then raised, thermal equilibration established, and the level read again. No correction is made for the thermal gradient between the bath and the portion of the capillary at room temperature. The bath is ultimately maintained constant at some predetermined value for periods up to 50 hours, and the meniscus level read as a function of time. The temperature of the bath is again raised, at intervals as before, up to a maximum about 100°C. The bath is then cooled in a manner similar to the heating cycle, until the system is at room temperature.

Data from a typical run are shown plotted in Fig. 2. The first portion of the warming, *AC*, up to the constant temperature, shows the characteristics of simple thermal expansion. The second portion, *CE, at constant temperature*, shows that the level decreases, or a net decrease in volume exists. Sufficient time may be allowed for all the possible change in volume to occur, but usually, at slower rates, the volume change is completed at more elevated temperature wherein the rate is faster. Further heating after the volume change is complete, and subsequent cooling and heating traces a reproducible path, *FH*. The cooling relationship is linear, but of different slope than the initial heating relationship. The difference in slope is indicative that the final material has a different coefficient of thermal expansion than the original. No attempt has been made to measure either coefficient.

Two quantities are evaluated from the dilatometer data. (1) S_x, defined as the extent of shrinkage, or loss in volume between initial and final states at 30°C. (2) $s^t h$, defined as the time when half the shrinkage is complete, at the temperature of measurement.

S_x, for a particular PVC with DOP, is 0.053±0.003 cc/cc of PVC. The second quantity, $s^t h$, is more easily visualized if the data are presented as in Fig. 2B. The shrinkage at the particular temperature of test is shown as "relative height," on which a value of unity represents no shrinkage, or the original thermal expansion line, and the value of zero, the final expansion and contraction line. On this presentation, $s^t h$ is the time when relative height is one-half.

The curves, so expressed, are reproducible within a few percent. The shape of the shrinkage curve has not been derived in any satisfactory manner. The slopes or half-times vary considerably with temperature, and may be plotted in the conventional Arrhenius manner, as in Fig. 3, evaluating an activation energy,

$$\Delta E^{**} = 80 \text{ kcal./mole.}$$

III. DIELECTRIC CONVERSION

An electrical method exists for following changes which occur in a plasticizing system. This method involves following the change in over-all dielectric constant of the system, as the change proceeds isothermally. Note that the dielectric constant[1] of plasti-

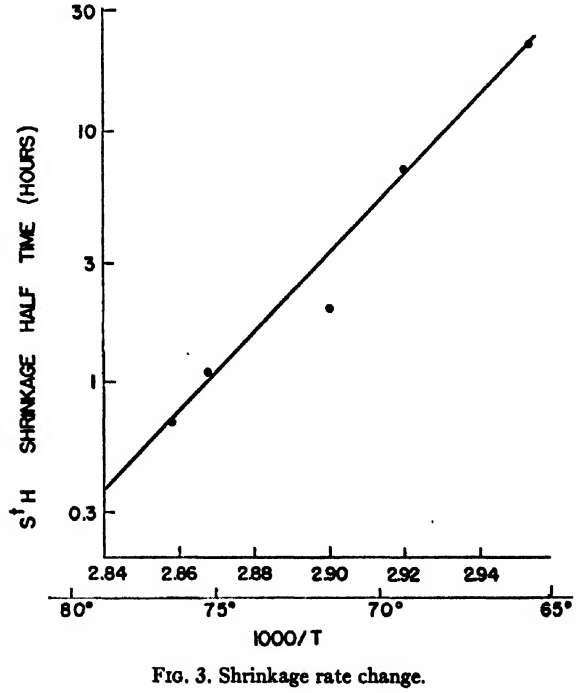


FIG. 3. Shrinkage rate change.

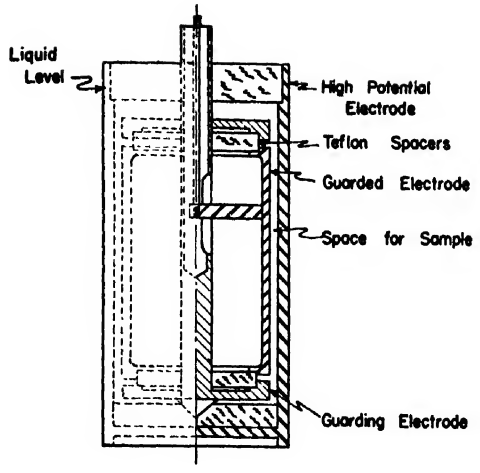


FIG. 4. Guarded cylindrical dielectric cell

cized PVC is greater than the volume fraction summation of its components. Thus, conditions may be established such that this change in dielectric constant may be evaluated in detail. These conditions are established, and measurements made, by procedures generally labeled "dielectric rate studies."

The material studied is a mixture of PVC and plasticizer, proportioned such that the mixture is slightly oily and easily manipulated. Plasticizer ratios are commonly 80–95 PHR, but have been varied from 20 to 150 PHR. The mixture is loaded into a three-electrode, guarded cylindrical cell, shown in Fig. 4.

The cell is placed in an oil bath, controllable to 0.05°C. About ¾ hour is allowed for thermal equilibration, then the (over-all) dielectric constant and loss factor are evaluated, using a Schering bridge at audio-frequencies. In the original state, the measured dielectric constant at any particular frequency is very close to the volume fraction summation of the components, if due allowance is made for entrapped air. This relationship is a limiting case of the Bruggeman formula for small spheres in an inert fluid.[5]

The temperature of the bath is raised about 10°C, thermal equilibration established, and dielectric measurements made again. This procedure is repeated until the temperature reaches some predetermined value. At this constant temperature, measurements are made as a function of time, with intervals ranging from a half-hour to a day each. The measured dielectric constant changes from the value associated with the original mix to a value approaching that expected for plasticized PVC having the same space factor.

The change in dielectric constant is not uniform, but is rapid at first, becoming slower as time progresses. The manner of change is similar to the exponential decay of a source material into a product.

$$\epsilon_t = \epsilon_1(e^{-mt}) + \epsilon_2(1-e^{-mt}), \qquad (1)$$

where ϵ_t = dielectric constant of the mixture at time, t
ϵ_1 = dielectric constant of the mixture originally
ϵ_2 = dielectric constant of the mixture finally.

Equation (1) may be re-expressed.

$$\ln(\epsilon_2-\epsilon_1)/(\epsilon_2-\epsilon_t) = mt. \qquad (2)$$

This is usually plotted

$$\log\Delta\epsilon/\Delta\epsilon_t = m't, \quad \text{as shown in Fig. 5.} \qquad (3)$$

Define a half-life, $_dt_h$, a time such that the fraction doubles, or the logarithm increases by ln2. This half-life is the inverse of the slope of the lines shown, and is a good measure of (relative) dielectric change rate. The half-life is temperature dependent, and a conventional Arrhenius plot, such as Fig. 6, permits an evaluation of the activation energy.

$$\Delta E^{**} = 110 \text{ kcal./mole.}$$

The change in dielectric constant noted during thermal equilibration is only one of several physical changes occurring. The mixture has changed from an oily slurry to a dry, crumbly, non-coherent powder. This powder is believed identical with the material known as "preplasticized PVC."

If this changed material is the same as fully plasticized PVC the change in dielectric constant should be accompanied by appearance of a loss factor peak near room temperature.[1] A "cooling curve" taken on this material showed the required peak. It was noted that the loss factor of the material increased with a decrease in temperature, probably the most characteristic behavior of polar materials near an "absorption" peak.

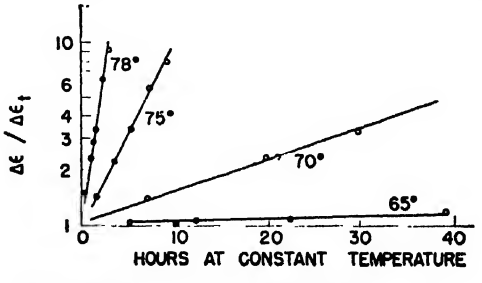


FIG. 5. Dielectric change in PVC-DOP systems.

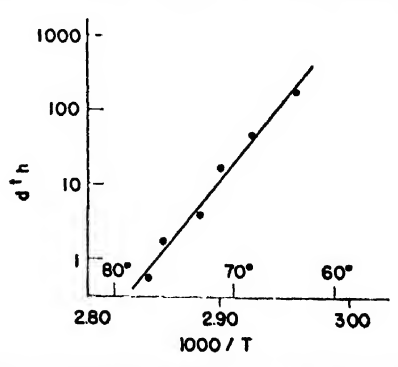


FIG. 6. Rate of change of dielectric properties with temperature.

[5] D. A. G. Bruggeman, Ann. d. Phys 24, 636–64 (1935).

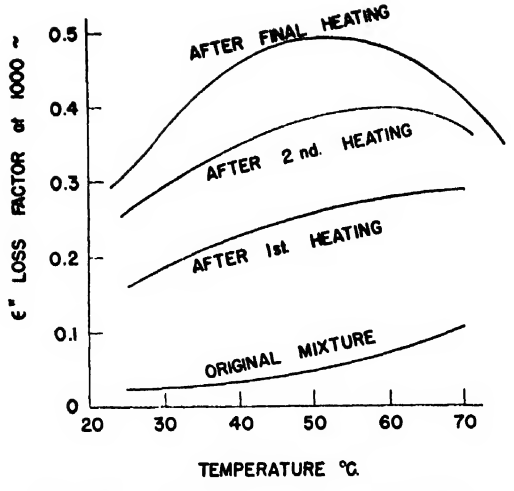


FIG. 7. Change of dipole peak during conversion.

IV. ASSOCIATED EXPERIMENTS

The dielectric data, as expressed by Eq. (2), imply that the final material is formed, as such, without any intermediate "dilution" steps or processes. This implication seemed so radical that other experiments were run to check this hypothesis. If the final material were so formed, it would be expected that the associated dipole peak would "grow" at a given temperature, for several investigators[1,3] have shown that the temperature of the loss factor peak depends upon the amount of plasticizer associated with the PVC. Hence, an experiment was run to try to identify dipole peaks at intermediate stages of this change, i.e., an interrupted conversion run.

The dielectric properties of a mixture in the cell were measured while the cell was warmed. Conversion was allowed to proceed about one $_{d}t_{h}$ only. The mixture was cooled quickly, and a similar warm-up curve obtained. At the same constant temperature, equilibration was established for an additional $_{d}t_{h}$. Quick cooling was followed by the measurement of another warm-up curve. Final heating to complete conversion, and subsequent cooling measurements gave a final curve. The data are presented in Fig. 7. Note that the loss factor peak changes in thermal location, in contrast to concepts of "growing at one temperature."

Another experiment, contributing interesting data, referred to as the "dilution run" was performed as follows. A mixture of PVC and DOP of lower plasticizer content than previously studied was heat-treated to assure complete conversion, but no fusion, i.e., preplasticized PVC was made. This material was then mixed with further DOP and thermal-dielectric curves obtained. The data are presented graphically in Fig. 8.

Note that the dipole peak in the warming portion of the cycle is that for the original plasticized PVC. The dipole peak after additional heat treatment is that for the final mixture. Similarly, note that the dielectric constant of any one mixture increased during the heat treatment. The low absolute values of dielectric constant are due to the low space factor, i.e., the large amount of entrapped air, for which no correction was made. It is concluded that the plasticized PVC of low plasticizer contents "dilutes" further in the same manner as the PVC itself.

V. DISCUSSION

It is important to cite reasons and facts for believing that the shrinkage and dielectric phenomena are two separate "reactions," even though time related. (1) The time sequences are in a fixed order, i.e., shrinkage is essentially complete before dielectric conversion occurs. Note that at 65°C, for example, $_{s}t_{h} \approx 25$ hours, $_{d}t_{h} \approx 200$ hours. (2) The activation energies of shrinkage and dielectric conversion are different. (3) The shrinkage data cannot be expressed by the same relationships that "linearize" the dielectric data.

While "shrinkage" seems to be a distinct step in the plasticization process, the exact meaning of the measurements taken is not clear. It is believed that shrinkage measures the loss in volume of the system as plasticizer enters the PVC particle. However, the data do not fit the relationships derived upon the basis of plasticizer diffusing into a solid spherical polymer particle, simultaneously undergoing a loss in volume.

The large change in dielectric constant noted durin

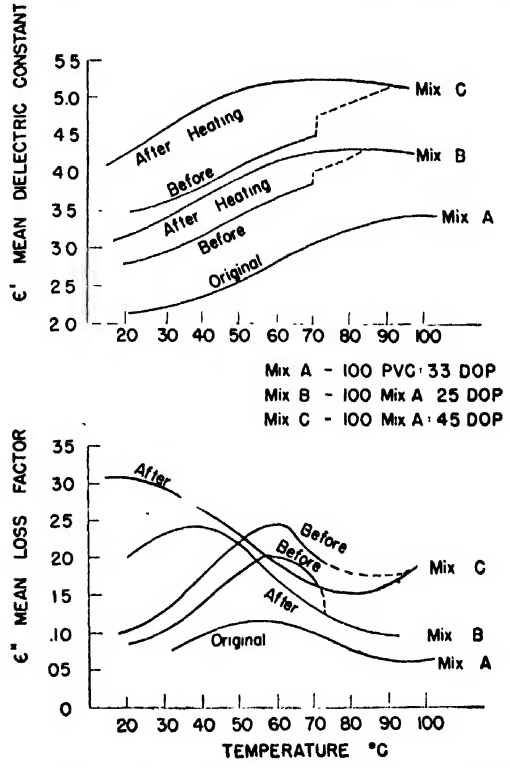


FIG. 8. Dielectric changes in dilution experiments. (A) Changes in measured dielectric constant. (B) Changes in measured loss factor.

plasticization is considered to be closely related to the other changes in physical properties. The assumption is made that these changes in properties are similar to those noted in simple liquids, as they change from solid to liquid phases. Thus, the increased dielectric constant of the plasticized PVC is due to the fact that part of the PVC is in its "liquid" form in which $\epsilon' \approx 11$, contrasted to $\epsilon' \approx 3$ in the "solid" form at room temperature. The "dilution data" suggest that a constant ratio of polymer may be converted from its "solid phase" to its "liquid phase," per unit of plasticizer, but that the plasticizer itself is unchanged in such a reaction. The concomitant effects of diluting the polymer in space allow the plasticizer to exhibit some of its own properties, such as color or viscosity, but such effects are secondary.

The changes in the real and complex part of the dielectric constant can be interpreted as follows. Discrete portions of the PVC chain, henceforth called "segments," have properties either "liquid" or "solid." While an individual segment may appear "liquid" or "solid," the composite behavior of two or more adjacent segments determines the loss factor. The segment is neither the polymerized monomer, nor the whole polymer molecule, but a portion of the chain about 10 vinyl groups or 20 carbon atoms long.

This segment concept of dielectric change leads to relationships identical with Eqs. (1), (2), and (3).

The segment is originally in the state wherein it shows a low dielectric constant. The change to the state wherein it shows the high dielectric constant follows rules of exponential decay, the rate constant depending upon plasticizer and temperature. Then, to a first approximation, the observed dielectric constant of a plasticized PVC system will be as in Eq. (4).

$$\epsilon = V_1\epsilon_1 + V_2\epsilon_2 \tag{4}$$

ϵ = experimentally observed dielectric constant
V_1 = volume fraction of PVC
V_2 = volume fraction of plasticizer
ϵ_1 = dielectric constant of PVC
ϵ_2 = dielectric constant of plasticizer.

In some cases ϵ_1 is apparently neither high nor low and Eq. (5) applies.

$$\epsilon_1 = V_\sigma\epsilon_\sigma + V_\lambda\epsilon_\lambda \tag{5}$$

ϵ_σ = low dielectric constant, about 3.5
ϵ_λ = high dielectric constant, about 11.5
V_σ, V_λ = relative fractions of each kind of PVC.

If an exponential decay process changes ϵ_σ to ϵ_λ, this may be expressed as

$$\epsilon_t = V_1[\epsilon_\sigma e^{-mt} + \epsilon_\lambda(1-e^{-mt})] + V_2\epsilon_2$$
$$\epsilon_t = V_1\epsilon_\sigma e^{-mt} + V_1(1-e^{-mt})\epsilon_\lambda + V_2\epsilon_2. \tag{6}$$

Then, using a procedure to derive a relationship similar to Eq. (2)

$$\ln\frac{V_1(\epsilon_\lambda - \epsilon_\sigma)}{(V_1\epsilon_\lambda + V_2\epsilon_2) - \epsilon_t} = ml. \tag{7}$$

Note that numerator of the fraction expresses the dielectric change in terms of concentration of PVC, and is independent of the kind of plasticizer. Similarly, the denominator expresses the difference between the final value and the instantaneous reading. Also note that the slope, m, or the half-life $_dl_h$, is independent of concentration, or of mix ratio of PVC and plasticizer.

The effect of plasticizer variation in the PVC-plasticizer system is expressed in a subsequent paper.[6]

VI. CONCLUSIONS

The plasticization of PVC by an ester such as DOP is a set of processes occurring in a definite sequence.

Two individual steps or processes have been identified, and labeled "shrinkage" and "dielectric conversion."

"Shrinkage" is a rate-process step not well understood.

"Dielectric conversion" is a rate-process step, which is independent of plasticizer-polymer mix ratio. The change in dielectric constant of a given mixture is related to the PVC content only. The change is assumed due to an unbinding of the PVC elementary segment, in the same manner normally accomplished by temperature alone.

VII. ACKNOWLEDGMENT

The author wishes to express his appreciation of the stimulating discussions with D. L. Loughborough concerning details of this work, and is indebted to the B. F. Goodrich Company for permission to publish the results of this investigation.

[6] M. L. Dannis, "Studies in plasticization of PVC. II. Plasticizer variations in PVC-plasticizer systems," J. App. Phys. **21**, 510 (1950).

Studies in Plasticization of Polyvinyl Chloride. II. Plasticizer Variations in PVC-Plasticizer Systems

M. L. Dannis
B. F. Goodrich Research Center, Brecksville, Ohio

Polyvinyl chloride resin is softened by certain liquids, called plasticizers, through a set of processes with high activation energies, 50–100 kcal. At fixed temperature, "reaction" rates are directly proportional to vapor pressure of the liquid. The dielectric change may be treated similarly to a chemical reaction between a solid and a vapor, and free energy of activation, ΔF^{**}, evaluated. ΔF^{**} is constant for various liquids with one resin. $\Delta E^{**}=\Delta H^{**}+L$ in this system, where ΔE^{**} is Arrhenius activation energy, ΔH^{**} energy barrier, L latent heat of vaporization of the liquid.

THE interaction of PVC and plasticizer (DOP) occurs in a complicated manner through a set of mechanisms apparently independent of concentration. Further information about these "reactions" was obtained in the study of the "reaction behavior" of PVC-plasticizer systems, varying the plasticizer composition as the major variable. The techniques and equipment used have been described previously.[1]

Since the introductory investigation had shown that the rate-curves and activation energy in both shrinkage and dielectric conversion were independent of mixing proportions, the mixes used in subsequent experiments were held constant within any one system, but varied as the plasticizer was changed, as determined by fluidity or compaction requirements.

The PVC used was a commercial polymer, Geon 101. Three different samples of resin were used, labeled *A*, *B*, and *G*. Samples *A* and *B* are believed to be identical. The plasticizers used were either commercial products of high purity, or laboratory prepared esters distilled in vacuum. All the liquids were used as received, with no further attempt at distillation or desiccation.

The systems studied as to shrinkage characteristics are listed subsequently in Table I. The data determining activation are shown in Fig. 1. The equations of the linear relationships shown, and the intercept with the 70°C ordinate, $_st_h$, are presented in Table I.

Note that in the systems with the plasticizers commonly considered the better "fluxing agents," the tendency is to "react" more easily. A fairly good comparison of relative shrinkage rates may be made by computing the shrinkage half-times at a particular temperature, since the slopes or activation energies are about the same magnitude. Note that the activation energies are relatively insensitive to molecular weight of the plasticizer, within the range $200<$M.W.<400. Note that the relative shrinkage rates, as measured by $_st_{70}$, are directly proportional to vapor pressure, as shown in Fig. 2. No other physical parameter of the plasticizer investigated, such as density, viscosity, molar volume, or surface tension gives a reasonable relationship.

The systems studied with respect to dielectric conversion characteristics are listed subsequently in Table II. The data for computing activation energies are shown in Fig. 3. The equations of the relationships shown are listed below in Table II, together with the $_dt_h$ value computed at 70°C, henceforth called $_dt_{70}$.

Note that in the systems with plasticizers commonly considered better "fluxing agents," the tendency is to convert more rapidly at a given temperature or the same speed at a lower temperature. Note that such "borderline" plasticizers as dilauryl phthalate and acetylated castor oil convert only slowly at elevated

TABLE I.

Shrinkage-rate equations:

$\log {}_dt_h = A_s[(1000/T)-B_s]$

$\ln {}_dt_h = \Delta E_s^{**}/RT - C_s/R$

$_dt_h$ = half-time in hours
T = absolute temperature
R = gas constant
ΔE_s^{**} = experimental activation energy cal./mole

System	A_s	B_s	ΔE_s^{**}	C_s	$_st_{70}$
PVC-A, DOP	15.9	2.866	72×10^3	208	5.8
B, TCP	15.0	2.910	69	200	1.15
B, DBP	12.4	3.003	56	170	0.085
PVC-G, DOP	14.5	2.852	66×10^3	189	7.8
G, DMT	14.7	2.888	67	196	2.4
G, TCP	15.2	2.896	70	202	1.9
G, B-16	13.0	2.942	60	175	0.42
G, DBP	11.4	2.990	52	156	0.133

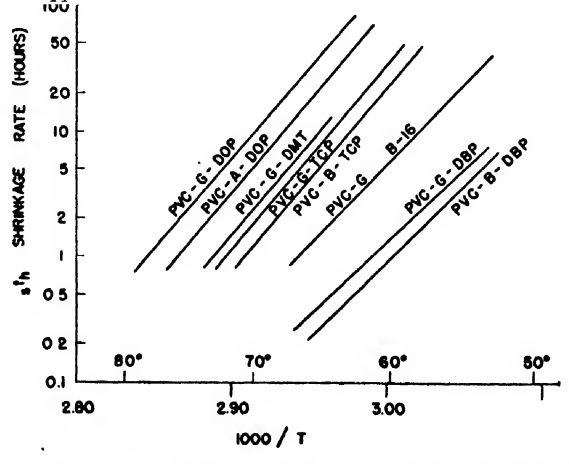


FIG. 1. Shrinkage rate characteristics in PVC-plasticizer systems.

[1] M. L. Dannis, "Studies in plasticization of PVC. I. Physical changes and their measurement," J. App. Phys. 31, 505 (1950).

temperatures, i.e., above 80°C, or above the second-order transition temperature of the PVC.

Note that ΔE_d^{**}, as a measure of polymer plasticizer interaction tendency, does not correlate with other factors commonly used, namely, "compatibility" in a non-solvent (hexane) dilution test, or the Huggins-Flory "μ" value, as evaluated by Doty and Zable.[2] Note however, that the d_{70} relationship is linear with vapor pressure, over four decades of the variables, as shown in Fig. 4. The vapor pressures used were taken from manufacturers' data, when possible.

THEORETICAL INTERPRETATION

Several facts have been presented about the dielectric conversion rates which classify this behavior as a new phenomenon. (1) The reacting system is two component, a liquid and a (solid) resin. (2) The initial reaction rate in any one system is independent of concentration or of mixing ratio. (3) The instantaneous reaction rate in any one system falls off in a simple exponential manner. (4) The reaction rate, in various systems at comparable conditions is proportional to the vapor pressure of the plasticizer.

The facts cited are consistent with the assumption of a reaction between a solid and a vapor. The concentration of plasticizer in the vapor state would be independent of the mix proportions of resin and plasticizer, but would depend greatly on the temperature. Similarly, the effective concentration of the solid would decay exponentially during the reaction, hence be the rate limiting step.

The dielectric conversion reaction has been expressed as Eq. (1) in the previous paper.[1]

$$\epsilon = \epsilon_1 e^{-mt} + \epsilon_2(1 - e^{-mt}). \tag{1}$$

The rate of change of dielectric constant, per unit change, is obtainable.

$$\frac{d\epsilon/(\epsilon_2 - \epsilon_1)}{dt} = me^{-mt}. \tag{2}$$

Impose the condition that t approaches zero,

$$\frac{d\epsilon/(\epsilon_2 - \epsilon_1)}{dt} = m. \tag{3}$$

So that m is the initial fractional rate of reaction of the system. This quantity has been evaluated experimentally, inasmuch as m is also the slope of the relationships presented previously in Fig. 5, of reference 1. Thus

$$m = \ln 2/3600 d_h. \tag{4}$$

From rate-process theory, as presented by Eyring, Glasstone *et al.*,[3,4] the "rate of reaction" of a system of two components is $C_A C_B k_r$, wherein C_A and C_B are the concentrations of the reactants, and k_r is the "specific reaction rate." k_r is defined

$$k_r = (kT/h)\exp(-\Delta F^{\circ **}/RT) \tag{5}$$

which is equivalent to the probability of finding a single activated complex in one cc of space, with a free energy of formation $\Delta F^{\circ **}$. Since actual experimental conditions do not involve the standard state, ΔF^{**} rather than $\Delta F^{\circ **}$, is the quantity discussed in further operations. Define the rate of reaction $= k_a$.

Then

$$k_a = C_A C_B k_r = m \tag{6}$$

if C_A and C_B are in appropriate units.

Since m has the dimensions of inverse time, C_A and

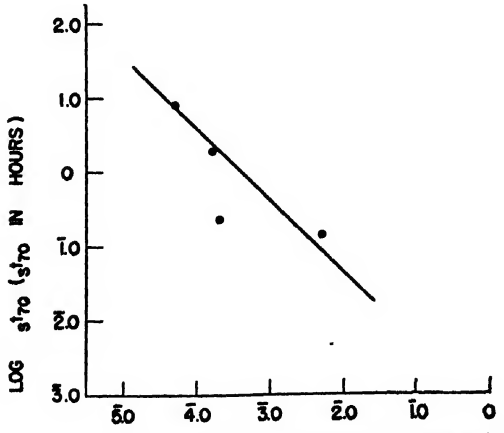


FIG. 2. Relationship of plasticizer vapor pressure to shrinkage rate.

TABLE II. Dielectric conversion characteristics.

$\log d_h = A_d[(1000/T) - B_d]$
$\ln d_h = \Delta E_d^{**}/RT - C_d/R$

d_h = dielectric half-time in hours
T = absolute temperature
R = gas constant
ΔE_d^{**} = experimental activation energy—cal./mole

System	A_d	B_d	ΔE_d^{**}	C_d	d_{70}
PVC-B-DnDP (Di-*n*-decyl phthalate)	19.0	2.807	87×10^3	244	110
PVC-A-DOP (Di-2-ethyl hexyl phthalate)	24.4	2.857	111	318	25
PVC-B-DnOP (Di-*n*-octyl phthalate)	22.4	2.862	103	294	15
PVC-B-TCP (*m*-tricresyl phosphate)	19.6	2.885	89	256	3.8
PVC-B-DMT (Dimethyl thianthrene)	22.7	2.899	104	302	2.3
PVC-B-B-16 (Butyl phthalyl butyl glycolate)	18.0	2.927	82	242	0.58
PVC-B-DBP (Dibutyl phthalate)	14.8	2.961	68	201	0.20
PVC-B-DEP (Di-ethyl phthalate)	17.6	3.020	80	243	0.0135
PVC-B-DMP (Dimethyl phthalate)	17.2	3.042	79	239	0.0063

[2] P. Doty and H. S. Zable, "Polymer-solvent interaction," J. Poly. Sci. 1, 90 (1946).

[3] Glasstone, Laidler, and Eyring, *Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941).

[4] S. Glasstone, *Textbook of Physical Chemistry* (D. Van Nostrand Company, Inc., New York, 1946), second edition.

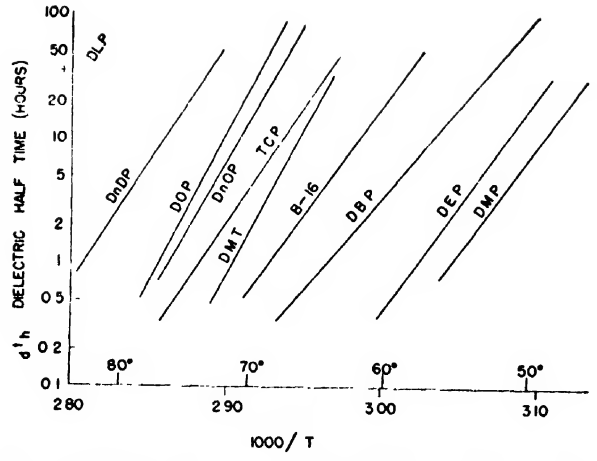


FIG. 3. Dielectric conversion rates in PVC-plasticizer systems.

C_B must be units of mole fractions. Evaluate ΔF^{**}

$$-\Delta F^{**} = RT[\ln(h/kT)+\ln(m/C_A C_B)]$$
$$\Delta F^{**} = -RT[\ln(h/kT)+\ln(\ln 2/3600 C_A C_B\, dl_h)]. \quad (7)$$

Since

$$\Delta F^{**} = \Delta H^{**} - T\Delta S^{**} \quad (8)$$

$$\Delta H^{**} - T\Delta S^{**} = -RT[\ln(h/kT) + \ln(\ln 2/3600 C_A C_B\, dl_h)]. \quad (9)$$

Evaluate (9) at T_1 and T_2, and subtract. Then

$$(-\Delta H^{**}/R)[(1/T_1)-(1/T_2)] = \ln(T_1/T)_2+\ln[(C_A C_B\, dl_h)_1/(C_A C_B\, dl_h)_2]. \quad (10)$$

But for small changes in temperature, $\ln(T_1/T_2)\approx 0$. C_A is the mole fraction of a solid, $(C_A)_2/(C_A)_1 \approx 1$. C_B is the mole fraction of plasticizer vapor.

$$C_B = \left(\frac{P}{760}\times\frac{273}{T}\right) \quad \text{or} \quad (C_B)_2/(C_B)_1 = \frac{(P/T)_2}{(P/T)_1}.$$

But the pressure of a vapor may be given by the Clapeyron equation

$$P_2 = P_1 \exp\{(L/R)[(1/T_1)-(1/T_2)]\}. \quad (11)$$

So

$$\ln[(C_B\, dl_h)_2/(C_B\, dl_h)_1] = \ln(T_1/T_2) - \ln[(dl_h)_2/(dl_h)_1] - (L/R)[(1/T_2)-(1/T_1)].$$

Thus

$$\ln[(dl_h)_2/(dl_h)_1] = (1/R)(-\Delta H^{**}+L)[(1/T_1)-(1/T_2)]. \quad (12)$$

Equation (12) is the method of evaluating ΔE^{**} in the usual Arrhenius plot of experimental activation energy. Thus

$$\Delta H^{**} = \Delta E^{**} - L - RT. \quad (13)$$

Note that Eq. (7) may be expressed as

$$\Delta F^{**} = \text{const} + \log P + \log dl_h. \quad (14)$$

This relationship is shown experimentally by Fig. 4. Figure 4 also implies that ΔF^{**} is a constant from one reacting system to another, or that the mechanism of plasticization is independent of the kind of plasticizer. However, ΔH^{**} and ΔS^{**} do vary from one system to another, depending greatly upon the choice of plasticizer. The values of ΔE^{**}, ΔH^{**}, and ΔS^{**} are presented in Table III. Assuming C_A = unity, $\Delta F^{**} = 16{,}500$ cal./mole at 70°C.

It is interesting to note that ΔS^{**} is the order of magnitude for dipole rotation of the PVC itself.[5]

No ordinary parameter of the molecule, such as molecular weight, molar volume, or viscosity seems to be related to these quantities in a simple manner.

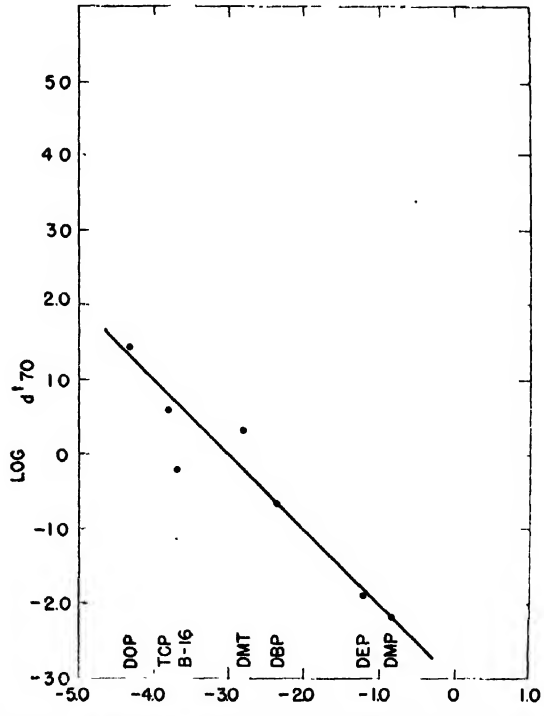
Some data recently obtained, have shown that ΔF^{**} is a variable depending upon the history of the preparation of the PVC.

CONCLUSIONS

The interaction of plasticizer with PVC is markedly dependent upon the vapor pressure of the plasticizer used, and largely independent of all other variables, except temperature.

One part of the interaction process may be described kinetically as the reaction of a solid and a vapor.

The free energy of formation of the activated complex in the critical rate limiting step in plasticization is a



LogP_{70} (P_{70} = vapor pressure of plasticizer) m/m Hg at 70°C

FIG. 4. Relationship of plasticizer vapor pressure to dielectric conversion.

[5] W. Kauzmann, "Dielectric phenomenon," Rev. Mod. Phys. 14, 16 (1942).

TABLE III. Thermodynamic constants of the activated complex in dielectric conversion.

System	ΔE^{**}	ΔH^{**}	ΔS^{**}
PVC-B, DnDP	87×10^3 cal./mole	59×10^3 cal./mole	125*EU*
A, DOP	111	83	196
B, DnOP	103	75	172
B, TCP	89	62	135
B, DMT	104	84	198
B, B-16	82	55	113
B, DBP	68	43	79
B, DEP	80	60	128
B, DMP	79	61	132

constant, regardless of kind of plasticizer used. This implies that the action of all plasticizers is fundamentally the same.

The Arrhenius activation energy, ΔE^{**}, obtained from a dielectric rate-plot, differs from the heat of activation, ΔH^{**}, by the latent heat of vaporization of the plasticizer used.

The interaction of plasticizer and PVC to form the "activated complex" leading to the plasticized material, does not depend on any simple, obvious property of the plasticizer, nor does it correlate with Huggins-Flory interaction coefficients previously evaluated for some of these systems.

ACKNOWLEDGMENTS

The author wishes to express his appreciation of the stimulating discussions with D. L. Loughborough concerning details of this work, and is indebted to the B. F. Goodrich Company for permission to publish the results of this investigation.

Mechanical Properties of Substances of High Molecular Weight. VIII. Dispersion of Dynamic Rigidity and Viscosity in Concentrated Polyvinyl Acetate Solutions*

JOHN D. FERRY, W. M. SAWYER,† GEORGE V. BROWNING,‡ AND ARTHUR H. GROTH, JR.
Department of Chemistry, University of Wisconsin, Madison, Wisconsin

The dynamic rigidities and viscosities of concentrated solutions of polyvinyl acetate in 1,2,3-trichloropropane have been obtained from transducer and wave propagation measurements. Two polymer samples with number-average molecular weights of 140,000 and 840,000 were studied in the lower audiofrequency range at concentrations from 3 to 40 percent and temperatures from −3 to 41°C. For each sample, when the reduced dynamic rigidity $G'T_0/Tc$ and the reduced dynamic viscosity η'/η were plotted against the reduced frequency, $\omega\eta T_0/Tc$, all the data superposed to give two composite dispersion functions. The distribution of Maxwellian relaxation times can be derived from the dispersion of either G' or η', and the two calculations are in reasonable agreement.

INTRODUCTION

STUDIES of wave propagation in concentrated polyvinyl acetate solutions have provided measurements of the wave rigidity, $\bar{G}$, in several different solvents at various frequencies, temperatures, and concentrations, as reported in the preceding paper of this series.[1] In one solvent, 1,2,3-trichloropropane, the wave damping can be measured with sufficient accuracy to calculate also the dynamic rigidity, G', and dynamic viscosity, η', which can be more directly interpreted. These data are now presented and are supplemented by direct measurements of G' and η' using transducer methods,[2] which are applicable at lower concentrations. When reduced to a standard viscoelastic reference state by a treatment derived elsewhere,[3] the results can be combined to provide dispersion functions for G' and η' which extend over several cycles of logarithmic frequency. The two functions are interrelated through a distribution function of relaxation times.

MATERIALS AND METHODS

The two samples of polyvinyl acetate employed in this work have been described elsewhere.[1,4] The number-average molecular weights of the unfractionated material and of Fraction 1 (a cut comprising 13 percent of the unfractionated) were 140,000 and 840,000, respectively.

Values of the wave rigidity, reported previously in graphical form,[1] were combined where possible with measurements of the damping index, λ/x_0, to give the dynamic rigidity and viscosity by the equations[5]

$$G' = \bar{G}\frac{4\pi^2[(4\pi^2-(\lambda/x_0)^2]}{[4\pi^2+(\lambda/x_0)^2]^2} \tag{1}$$

$$\eta' = \frac{\bar{G}}{\omega}\frac{16\pi^3\lambda/x_0}{[4\pi^2+(\lambda/x_0)^2]^2}. \tag{2}$$

* Presented at the Seventh Meeting of the Division of High-Polymer Physics, American Physical Society, New York City (February 2–4, 1950).

† Present address: Shell Development Company, Emeryville, California.

‡ Present address: Standard Oil Company, Whiting, Indiana.

[1] W. M. Sawyer and J. D. Ferry, J. Am. Chem. Soc. (to be published).

[2] Smith, Ferry, and Schremp, J. App. Phys. **20**, 144 (1949).

[3] J. D. Ferry, J. Am. Chem. Soc. (to be published).

[4] G. V. Browning and J. D. Ferry, J. Chem. Phys. **17**, 1107 (1949).

[5] Ferry, Sawyer, and Ashworth, J. Polymer Sci. **2**, 593 (1947).

TABLE I. Dynamic rigidity and viscosity of solutions of unfractionated polymer from wave propagation measurements.

Conc., wt. percent	Temp. °C	ν	$G' \times 10^{-4}$ dyne/cm²	η' poises
23.8	0.1	250	9.7	35.0
		320	11.5	31.1
		400	15.1	29.2
		500	13.4	23.8
		630	15.9	21.8
		800	15.3	18.5
		1000	15.8	14.8
	10.5	250	8.0	31.8
		320	10.9	31.6
		400	10.7	20.6
		500	12.1	19.0
		630	11.7	16.8
		800	12.3	13.0
		1000	12.9	14.1
35.1	0.0	500	43.8	82.3
		630	46.8	65.0
		800	51.6	51.4
		1000	54.6	44.2
		1600	57.4	29.0
		2000	52.7	24.6
	10.4	500	35.5	68.5
		630	40.4	59.2
		800	40.8	38.6
		1000	43.0	29.7
		1250	47.6	26.7
		1600	49.7	24.0
	30.3	400	23.3	53.3
		500	27.2	41.5
		630	27.3	32.6
		800	28.3	23.2
		1000	29.6	20.6
		1250	29.0	20.0

TABLE II. Dynamic rigidity and viscosity of solutions of Fraction I from wave propagation measurements.

Conc., wt. percent	Temp. °C	ν	$G' \times 10^{-4}$ dyne/cm²	η' poises
14.3	0°	1000	6.9	7.0
17.0	0.4	200	6.6	32.8
		250	7.9	23.4
		320	8.1	18.2
		400	6.8	11.4
		500	8.7	11.2
		630	10.4	11.0
		800	11.5	9.2
		1000	11.5	8.6
		1250	12.0	7.6
		1600	11.7	6.7
	14.9	250	6.9	24.5
		320	7.6	15.8
		400	7.2	13.5
		500	8.3	12.2
		630	7.6	9.5
		800	8.9	8.8
		1000	9.8	8.4
		1250	8.0	6.8
	25.6	250	5.6	16.3
		320	6.6	12.6
		400	6.9	10.4
		500	6.8	8.3
		630	7.6	8.0
		800	7.6	7.0
		1000	8.3	6.0
		1250	8.1	5.7

Transducer measurements were made with the Model 2 transducer of Smith, Ferry, and Schremp.[2] The original experimental arrangement was improved by mounting the transducer in a box of sawdust, carefully shielding all leads, and replacing the a.c. operated decade amplifier by a Ballantine[6] Model 220 battery operated amplifier. An isolating transformer was substituted for the amplifier previously employed between the oscillator and impedance bridge. The bridge was calibrated. The mechanical resistance and reactance were calculated by the circle plot method, and the dynamic viscosity and rigidity were obtained from plots of these quantities against depth of immersion of the oscillating rod.[2]

EXPERIMENTAL RESULTS

The results of the wave propagation measurements on the two samples are given in Tables I and II, and those of the transducer measurements in Tables III and IV.

It has been shown[3] that if certain simple postulates regarding stress relaxation in concentrated polymer solutions are valid, dynamic data at various frequencies, temperatures, and concentrations should superpose to form two universal dispersion functions when the reduced dynamic rigidity, $G_r' = G'T_0/Tc$, and the reduced dynamic viscosity, $\eta_r' = \eta'/\eta$, are plotted against the reduced frequency, $\omega_r = \omega\eta T_0/Tc$. Here η is the steady flow viscosity (at small shearing stresses), c the concentration in g polymer per cc solution, T the absolute temperature, and T_0 a standard temperature, taken as 298°K. From the data of Tables I–IV, together with steady flow viscosities measured on these same solutions,[7] and values of c computed by assuming additivity of volumes of polymer and solvent,[1] the quantities G_r', η_r', and ω_r have been calculated and are plotted logarithmically for the two samples in Figs. 1 and 2. In each case the open circles are transducer data and the solid circles are wave propagation data from individual damping measurements. The heavy dashed curves for G_r' are obtained by applying a constant correction to the previously published composite curve for reduced dynamic rigidity,[1] $\tilde{G}_r = \tilde{G}T_0/Tc$; the ratio $G'/\tilde{G}$ is fairly insensitive to the damping index when $\lambda/x_0 < 2\pi$, as shown by Eq. (1), and the variation of λ/x_0 is small enough over the ranges of frequency and concentration employed in the wave measurements so that a fairly good approximation is achieved by setting $G_r' = 0.90\tilde{G}_r$.

The results of the two entirely different experimental methods fit together very well. By the use of reduced variables, the data for unfractionated polymer covering a range of concentrations from 5 to 40 percent and of

[6] Ballantine Laboratories, Boonton, New Jersey.

[7] Ferry, Browning, Sawyer, and Foster, J. Phys. Coll. Chem. (to be submitted).

TABLE III. Dynamic rigidity and viscosity of solutions of unfractionated polymer from transducer measurements.

Conc., wt. percent	Temp. °C	ν	$G' \times 10^{-4}$ dyn/cm²	η' poises
5.00	25.0	200	—	1.02
		320	—	1.20
		400	—	1.15
10.15	25.0	200	0.266	5.26
		250	0.318	4.62
		320	0.381	3.79
		400	—	3.69
	35.0	200	0.147	4.74
		250	0.255	3.88
		320	0.309	3.45
		400	—	3.15
15.8	25.0	200	1.40	11.2
		250	1.75	10.4
		320	1.98	9.8
		400	1.91	8.4
	35.0	200	—	12.9
		250	-	10.1
		320	1.34	9.1
		400	1.53	8.3
20.0	25.0	200	3.49	25.1
		250	3.90	20.4
		320	3.98	16.8
		400	4.19	15.2
24.1	25.0	200	7.25	35.6
		250	7.27	32.6
		320	7.50	24.2
		400	7.26	21.2
	35.0	200	5.25	31.8
		250	—	27.1
		320	6.60	22.8
		400	7.39	20.4

TABLE IV. Dynamic rigidity and viscosity of solutions of Fraction I from transducer measurements.

Conc., wt. percent	Temp. °C	ν	$G' \times 10^{-4}$ dyne/cm²	η' poises
3.0	25.0	200	- -	1.20
		250	—	1.08
4.0	25.0	200	—	2.20
		250	—	1.74
		320	0.13	1.48
		400	0.22	1.45
5.0	25.0	200	—	3.02
		250	—	2.29
		320	—	1.95
		400	—	1.83
5.54	25.0	250	0.21	3.31
		320	0.23	2.71
		400	0.29	2.40
8.21	25.0	200	—	7.58
		250	0.76	6.41
		320	0.84	4.55
		400	0.85	4.10
11.3	25.0	250	1.78	8.70
		320	2.17	7.91
		400	2.28	6.48
14.35	25.0	250	3.95	13.1
		320	4.54	11.9
		400	4.81	10.2

temperatures from −3° to 41°C superpose rather closely to give two universal dispersion curves, and the data for Fraction I superpose similarly over a somewhat smaller range of temperatures and concentrations. As the frequency increases by 10^5, the rigidity increases by about 10^2, while the viscosity falls off by about 10^{-4}.

The functions as represented in the figures are reduced to a hypothetical viscoelastic reference state of unit density (concentration) and unit steady flow viscosity (1 poise) at 298°K. Although this state is of course non-existent, it provides a convenient basis for combining data, and approximate dispersion curves for any real state of known temperature, concentration, and steady flow viscosity should be obtainable from Fig. 1 or 2 by adjusting the coordinates. If the postulates[3] that all relaxation mechanisms depend identically on temperature and concentration do not hold strictly, some distortion of the dispersion functions may arise from combining data in this way. However, here as in previous examples[3] the superposition is successful insofar as the data at different concentrations and temperatures overlap, and it is believed that any such distortion is slight.

It is of interest also to plot $\omega_r\eta_r'$ (which is the imaginary part of the reduced complex dynamic rigidity,[5] $G_r''=G''T_0/Tc$) and the mechanical loss factor, $G_r''/G_r'=G''/G'$, against the reduced frequency. These functions are shown in Fig. 3. In both samples G'', like G', increases rapidly with increasing frequency. Their ratio, however, falls off with increasing frequency and appears to approach a constant value of about 0.5.

An attempt to apply the complex plane plot of Cole and Cole[8] showed that the locus of $G'/(G'^2+G''^2)$ and $G''/(G'^2+G''^2)$ is not a circle.

THE RELAXATION DISTRIBUTION FUNCTION

The dispersion curves for G' and η' are not independent; a formulation of the relation between them can provide an additional experimental check. This is most conveniently written in terms of a distribution function of Maxwellian relaxation times, $F(\tau)$, such that each elastic contribution to the rigidity $F(\tau)d\tau$ is relaxed as though by a dashpot $\tau F(\tau)d\tau$.[9] The exact equations for G' and η' in terms of $F(\tau)$ are[10,11]

$$G'(\omega)=\int_0^\infty \frac{F(\tau)\omega^2\tau^2}{1+\omega^2\tau^2}d\tau; \quad \eta'(\omega)=\int_0^\infty \frac{\tau F(\tau)}{1+\omega^2\tau^2}d\tau. \quad (3)$$

[8] K. S. Cole and R. H. Cole, J. Chem. Phys. **9**, 341 (1941); A. W. Nolle, J. Polymer Sci. **5**, 1 (1950).

[9] Although a set of finite mechanisms is more convenient for discussing the effects of temperature and concentration on dispersion (see reference 3), the equivalent continuous distribution is more useful here.

[10] T. Alfrey and P. Doty, J. App. Phys. **16**, 700 (1945).

[11] Kuhn, Künzle, and Preissmann, Helv. Chim. Acta **30**, 307 (1947).

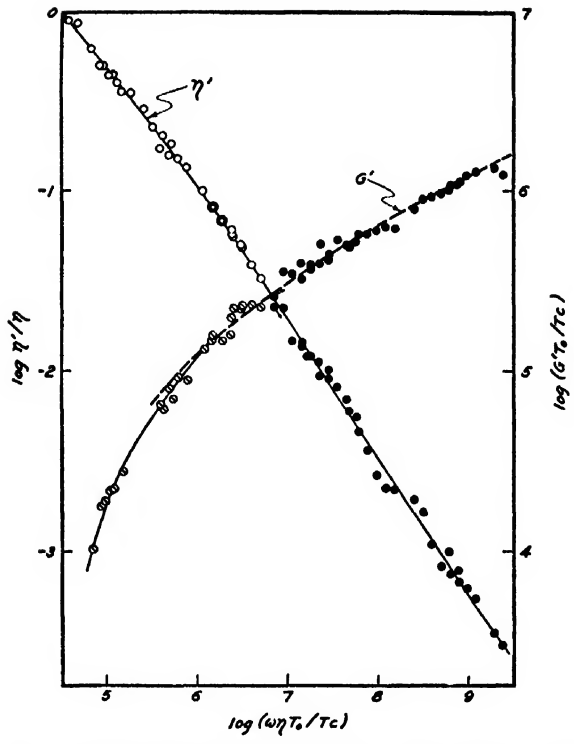


FIG. 1. Reduced dynamic rigidity and viscosity plotted against reduced frequency, for unfractionated sample. ○, η' from transducer measurements; ⊘, G' from transducer measurements; ●, η' and G' from wave propagation measurements with individual damping determinations; · · ·, displaced composite curve for $\tilde{G}_r$, covering concentration range from 16 percent to 40 percent polymer (reference 1, Fig. 3), corresponding to the approximation $G_r' = 0.90\tilde{G}_r$.

For a broad distribution we may use an approximation,

$$G'(\omega) = \int_{1/a\omega}^{\infty} F(\tau)d\tau; \quad \eta'(\omega) = \int_{0}^{1/a\omega} \tau F(\tau)d\tau \qquad (4)$$

such that those Maxwell elements with τ above a certain cut-off time $1/a\omega$ contribute only to the rigidity while those below contribute only to the viscosity. In the use of this approximation by Guth,[12] the numerical factor a was taken as unity; we must make it slightly different from unity to correlate the experimental values of G' and η'.

We define a logarithmic distribution function $\Phi(\ln\tau)d\ln\tau = F(\tau)d\tau$ and differentiate the definite integrals of Eqs. (4) with respect to their limits to obtain $dG'/d\ln\omega = \Phi(-\ln a\omega)$; $d\eta'/d\ln\omega = -(1/a\omega)\Phi(-\ln a\omega)$. For calculations from logarithmic plots such as Figs. 1 and 2, it is convenient to solve for Φ as follows:

$$\Phi(-\ln a\omega) = G'(d\log G'/d\log\omega) = -a\omega\eta'(d\log\eta'/d\log\omega). \qquad (5)$$

In terms of reduced variables, if Φ_r is defined as $\Phi T_0/Tc$, we have

$$\Phi_r(-\ln a\omega_r) = G_r'(d\log G_r'/d\log\omega_r) = -a\omega_r\eta_r'(d\log\eta_r'/d\log\omega_r). \qquad (6)$$

Thus Φ_r may be calculated either from the dispersion of G_r' or the dispersion of η_r'. Excellent agreement between the two methods is obtained for the unfractionated sample if a is taken as 0.77, as shown in Fig. 4, where Φ_r is plotted against $\log\tau_r$ (here $\tau_r = \tau cT/T_0\eta$). The same value of a applied to the data for Fraction I gives agreement at low τ_r but some divergence at high τ_r (Fig. 5). At high τ_r, the values from η' are considered to be more reliable.

The figures also show curves for $\tau_r\Phi_r$. The values of Φ_r and $\tau_r\Phi_r$ approach zero at high and low values of the abscissa, respectively, so that the integrals of Eqs.

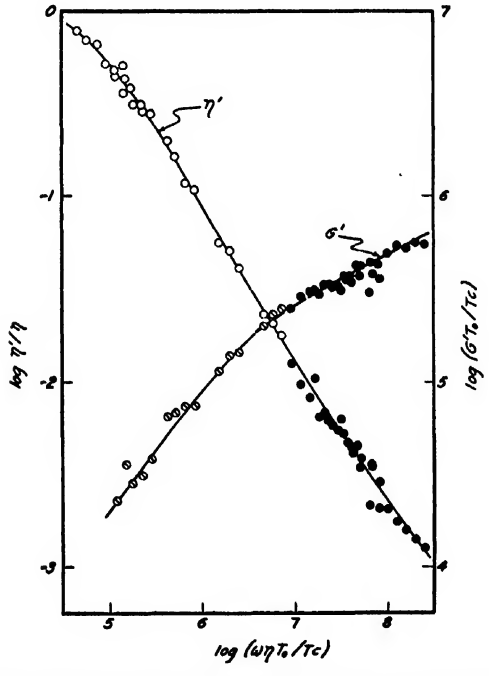


FIG. 2. Reduced dynamic rigidity and viscosity plotted against reduced frequency, for Fraction I. Key to symbols same as in Fig. 1. The displaced composite curve for $\tilde{G}_r$ (reference 1, Fig. 8) includes concentrations of 14 percent and 17 percent polymer.

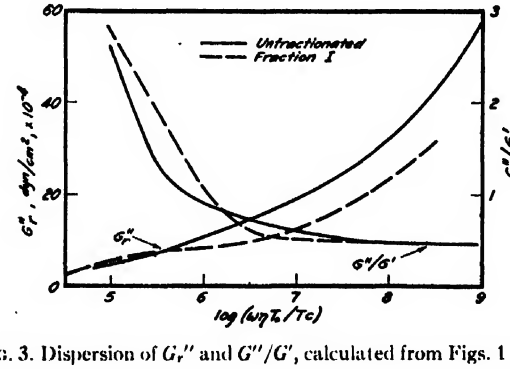


FIG. 3. Dispersion of G_r'' and G''/G', calculated from Figs. 1 and 2.

[12] Ivey, Mrowca, and Guth, J. App. Phys. 20, 486 (1949).

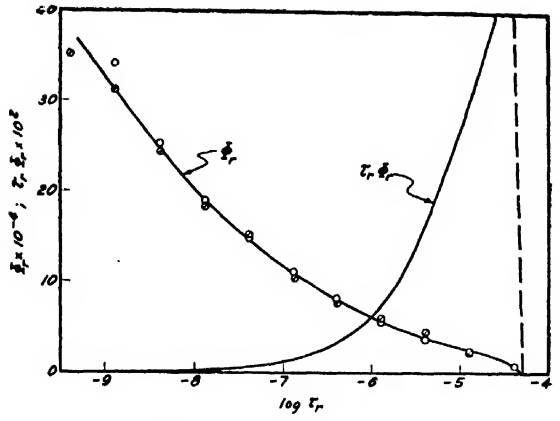


FIG. 4. Distribution functions Φ_r and $\tau_r\Phi_r$ for unfractionated sample, calculated from dispersion of $\eta'(\circ)$ and dispersion of $G'(\varnothing)$.

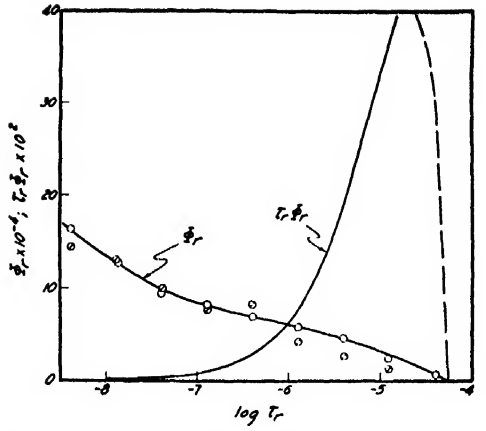


FIG. 5. Distribution functions Φ_r and $\tau_r\Phi_r$ for Fraction I. Key to symbols same as in Fig. 4.

(4) are convergent and may be used to calculate G_r' and η_r' at any desired frequency. Such calculations check the experimental values within graphical errors. The integral of $\tau_r\Phi_r d\ln\tau_r$ from $-\infty$ to ∞ must be unity because of the definitions of the reduced variables, and this is found graphically to be the case. Alternatively, G_r' and η_r' can be calculated from Φ_r by graphical integration of the exact Eqs. (3). These calculations check within about 10 percent, reflecting probably the degree of accuracy of the cut-off approximation. Thus a single empirical function Φ, together with data on the steady flow viscosity η, can provide an approximate description of the dynamic mechanical properties in small deformations over a wide range of frequencies, temperatures, and concentrations.

For the high molecular weight sample (Fig. 5), the function Φ is of the order of 10^6 dyne/cm^2 and, over several decades, has a relatively low slope. It is probable that the existence of such a plateau-like region in the relaxation function of high molecular weight polyisobutylene is responsible for the partial success of the box distribution approximation[11] in fitting stress relaxation data on this polymer.[13] For the low molecular weight polyvinyl acetate (Fig. 4), the slope of Φ is steeper, and it is not expected that the box approximation would be applicable to such a material. It would be of interest to obtain an extension of Φ in the direction of lower τ from high frequency measurements such as those of Mason and Baker.[14] The detailed shape of the function at the high τ end should be obtainable from stress relaxation measurements in concentrated solutions.[15]

ACKNOWLEDGMENT

This work was supported in part by the Research Committee of the Graduate School of the University of Wisconsin from funds supplied by the Wisconsin Alumni Research Foundation, and in part by a grant from Research Corporation.

[13] Andrews, Hofman-Bang, and Tobolsky, J. Polymer Sci. **3**, 669 (1948).

[14] Mason, Baker, McSkimin, and Heiss, Phys. Rev. **73**, 1074 (1948).

[15] F. W. Schremp and J. D. Ferry (unpublished work).

Stress Phenomena from the Respective Viewpoints of Solid-State and High Polymer Physics

MAURICE L. HUGGINS
Research Laboratory, Eastman Kodak Company, Rochester, New York

Flow and elastic properties are considered, in a qualitative manner, for the following types of materials: crystalline solids, simple liquids, glasses, and high polymers.

BY "stress phenomena" we mean here certain phenomena associated with the displacements of atoms or molecules relative to their neighbors when acted on by a shearing force or "shear stress." We are interested in the following: (1) The permanence of the displacements: On this basis we distinguish between elasticity, and flow or fracture. (2) The magnitude of the displacements, as related to the magnitude of the stress: Elastic moduli and stress-strain curves, for example, are expressions of this relationship. (3) The rate of displacement, as dependent on the stress: This is usually expressed by the fluidity or its reciprocal, the viscosity.

We are interested in the interpretation of these properties in terms of atomic and molecular structures and interatomic and intermolecular forces. Since the type of structure varies greatly in different classes of materials—crystals, liquids, glasses, high polymers, etc. —it is convenient to consider these classes separately.

In *crystalline solids* the chief effect of an applied stress is to shift all of the atoms or molecules slightly from their equilibrium positions in the crystal structure. Work is done against the interatomic or intermolecular forces; the energy of the system is increased. On removal of the stress, the atoms and molecules return to their equilibrium positions. The substance exhibits *short-range elasticity.*

If *flow* occurs at all, it takes place in jumps, one atom or molecule at a time. The structure after each jump is closely similar to that before the jump; hence the shifts are permanent and this is flow rather than elasticity. The jumps occur readily only where there are irregularities—departures from the perfectly crystalline arrangement—as at grain boundaries or regions where, for accidental reasons, there happens to be either a deficiency or an excess of atoms or molecules (per unit of volume) over the number required for the perfect crystal. Because of the crystalline arrangement, the shifts occur preferentially in certain directions, parallel to the atomic or molecular planes; hence the flow properties of the whole vary with direction.

To break or *fracture* a crystalline solid, the applied stress, greater than a certain "yield value," produces a shift of an atom or molecule to a less stable situation (e.g., one involving greater atomic separations); this shift facilitates a similar shift of a neighbor, such shifts then occurring autocatalytically across the crystal. This process is obviously also directional, the breaks tending to follow the planes of easiest cleavage in the crystal.

In simple *liquids*, irregularities of structure are numerous and irregularly distributed throughout the occupied volume. At some of these irregularities, shifts are constantly (or at least frequently) occurring, as a result of the temperature motions of the molecules. Without the application of an external stress, there is—within the errors of macroscopic observation—as much motion in any one direction as in the opposite direction; hence no over-all motion is observed. Application of a shear stress favors motions having a component in the direction of the stress and hinders those having a component in the opposite direction; *flow* is therefore observed. The rate of flow (hence the fluidity and its reciprocal, the viscosity) depends on the difference between the rate at which energy humps restricting molecular shifts are passed over in the forward and backward directions, and so on the magnitudes of these humps, their number, the temperature, and the applied stress. Unlike the crystalline case, the flow properties are independent of direction. For pure shear stress, there is (in a simple liquid) no "yield value"—an infinitesimal stress is sufficient to produce flow. *Elasticity* (short range) is observed for over-all compression, but not for shear.

Glasses are like liquids in having an irregular structure, but like crystalline solids in having (at ordinary temperatures) few or no atomic or molecular shifts occurring spontaneously—as a result of the temperature motions —or on the application of small shear stresses. Thus they exhibit *short-range elasticity*, and *flow* only with high stresses, long times, or high temperatures. The detailed interpretation of stress phenomena is complicated, for ordinary silicate glasses at least, by the fact that the structures contain charged ions, these being of a variety of chemical types, charges, sizes, and shapes. The existence of a substance in the glassy "state" is a result of the fact that the component atoms can be distributed in space in various ways of about equal stability, but with high energy humps hindering transitions from each such way to others.

High polymers are generally, like glasses, irregular in structure, although they frequently have regions which are more or less crystalline. The crystalline arrangement and many non-crystalline arrangements are usually of nearly equal energy. Moreover, the shifts of atoms needed to change a small non-crystalline region into a

crystalline one often do not require the breaking of any strong bonds—passage over large energy humps. On the other hand, to change a sample of a semicrystalline linear high polymer, in which many of the polymer chains run through two or more crystalline regions as well as amorphous intercrystalline regions, into a single crystal would require either the breaking of many strong bonds or the breaking up of many of the crystalline regions. This would involve passage over large energy humps, so large that the process would not occur at ordinary temperatures; at high temperatures the crystalline arrangement would not be stable relative to an amorphous structure. For these reasons, single crystals of linear high polymers have not been obtained.

Rotation around a single bond in an organic molecule, such as a linear high polymer, ordinarily involves only small changes in energy, of the order of magnitude of the average energy of temperature motion of the atoms. The polymer molecules can therefore change their shapes and their steric relations to their neighbors by motions of a few atoms at a time over relatively small energy humps. Both *flow* and *long-range elasticity* involve such changes. For flow, molecules must permanently shift their positions (e.g., their centers of gravity) relative to their neighbors—more of them shifting in the direction favored by the applied stress than in the opposite direction. In high polymers this shift rarely takes place all at once; it occurs, rather, as a succession of small shifts of portions of the molecule. The stress must be maintained over a time long enough for these successive shifts, and allowing for occasional shifts in the reverse direction—against the stress—in accord with the principles of equilibrium.

A single crosslink to an adjacent molecule can prevent the flow of a given molecule. The molecule can then change its shape, extending itself in the direction favored by the stress, but, on release of the stress, it will tend to revert to its original state, relative to its neighbors. This reversion may not be exactly achieved for a given molecule, but the average over all molecules of this sort will be such that there will be no permanent over-all shift, either in the forward direction or in the opposite direction. This reversible change of molecular shape is responsible for long-range elasticity.

The relative flow of adjacent molecules can be hindered or prevented by any intermolecular interactions of sufficient strength, such as hydrogen bonds or strong polar attractions. The cooperative action of many small forces is also effective. This occurs, for example, in crystalline regions: A portion of a chain molecule cannot be removed from such a region one segment at a time, a very large force being required to pull many segments simultaneously over their small energy humps. Even without crystallinity, a portion of a polymer chain which is extended as much as possible in the direction favored by the stress can move only as a unit. Motion of this unit obviously requires passage over a large energy hump—the summation of a series of small humps; it thus occurs less readily than motion of a smaller portion of the chain by itself.

Another factor to be considered is the extent to which a given shift of a segment in a polymer chain requires the pushing aside of neighboring regions, hence cooperating shifts in other molecules. This depends on the sizes, shapes and rigidities of the chain segments, including attached atoms or groups. Even the orientation, *cis* or *trans*, of the single bonds in a polymer chain adjacent to a chain double bond is important. With the *cis*-orientation (that in natural rubber), smaller segments may be shifted at a time and smaller volumes of the surroundings need to be rearranged, than with the *trans*-orientation (that in gutta percha). This is probably the explanation for the differences in long-range elasticity observed for natural rubber and gutta percha.

It is obviously possible for a material to exhibit both plastic flow and long-range elasticity at the same time. This is the case if, under the action of the applied stress, some of the molecules "flow" relative to their neighbors, while others merely change their shapes reversibly.

The forces restricting the motion of chain segments to give plastic flow and long-range elasticity are similar to those restricting the rotations of chains around the chain axes—the kind of motion which many types of chain molecules take up at their "second-order transition" temperatures, somewhat below the point at which they become fluid. The energy humps restricting both types of motion are similar in magnitude. The second-order transition point can therefore be used, for a given type of polymer at least, as a semiquantitative measure of its ability to undergo deformation of shape, either by plastic flow or by elastic extension, in the evaluation of plasticizers, for instance. For a better understanding of the plastic or elastic behavior of a substance or material, however, it is necessary to know something of the numbers of energy humps of different sizes—or, as some would prefer to express it, the distribution of "activation energies" or the "spectrum" of "relaxation times."

From the foregoing, it is evident that, although the same fundamental principles are involved in the different classes of substances considered, the application of these principles leads to widely different stress phenomena, as macroscopically observed.

Some Theoretical Considerations of Dynamic-Property Data on Textile Specimens

W. James Lyons
Chemical and Physical Research Laboratories, The Firestone Tire and Rubber Company, Akron, Ohio

1. Early stretch-vibrometer data, of quite extensive character, on the frequency dependence of the internal friction μ of a Nylon monofil, has been found to conform to the relation

$$\mu=[\mu_2/(1+\omega^2\tau_2^2)]+\mu_3$$

which was deduced independently by Tobolsky and Eyring from considerations of molecular structure. Among other constants evaluated was the relaxation time for the secondary-bond network τ_2, which was found to have the value 9.3 sec.

2. By more sensitive graphing methods than had been used heretofore, it was confirmed by the Nylon data that the hyperbolic relationship between resonant frequency ω_0 and internal friction μ holds equally as well as the foregoing equation. The two relationships are shown to be nearly equivalent in the present experimental range. On evaluation of graphical parameters, the equation $\mu\omega_0=2.47\times10^9$ ergs/cm³ was obtained.

3. Using the same graphical method with a set of data on 11/4/2 cotton cord the relation $(\mu+0.9)\omega_0=7.05\times10^9$ ergs/cm³ was established.

INTRODUCTION

In the application of a relaxation mechanism of the Kuhn type to dynamic experiments, Tobolsky and Eyring[1] indicate that under suitable conditions of load and frequency (in the range 20 to 200 c.p.s.) the modulus E and internal-friction coefficient μ, dynamically measured, are given by the equations

$$E=E_1+E_2, \tag{1}$$

and

$$\mu=[\mu_2/(1+\omega^2\tau_2^2)]+\mu_3, \tag{2}$$

where E_1 and E_2 are partial moduli governing the elastic deformation of the primary and secondary bonds, respectively; μ_2 and μ_3 represent friction, respectively, in the slippage of the secondary bonds, and in the motion of segments of the linear molecule relative to other parts of the chain; ω is the angular frequency of vibration, and τ_2 is the relaxation time for secondary-bond processes.

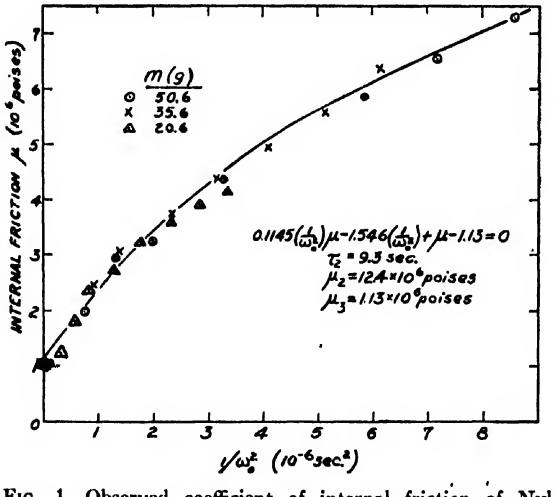


FIG. 1. Observed coefficient of internal friction of Nylon monofil as a function of $1/\omega_0$, at three vibrating masses m. Fitted to the data is a graph of Eq. (2).

The partial moduli E_1 and E_2 being assumed to be material constants in this theory, the dynamic parameter E, according to Eq. (1), should be independent of frequency. In line with its earlier confirmation for rubbers at frequencies above the very lowest,[2,3] this theoretical conclusion has been confirmed on the textile stretch-vibrometer for samples of nine different textile materials,[4,5] up to frequencies between 300 and 400 c.p.s. At frequencies approaching, and in the supersonic range, the frequency-independence of E for textiles probably disappears, as it does for other high polymers.

It has been of further interest to investigate the validity of the internal-friction relation, Eq. (2). For this purpose the extensive data previously obtained on a high tenacity Nylon monofil were selected for treatment.

NYLON MONOFIL DATA

The calculated coefficients μ, obtained on the stretch-vibrometer for eight different lengths of the Nylon monofil, with three vibrating masses (reference 5, Fig. 8), were plotted as functions of $1/\omega_0^2$, $\omega_0(=2\pi f_0)$ being the resonant angular frequency at which the particular measurement was made. The result is shown in Fig. 1.

It can be readily verified that Eq. (2) can be written as an equation in $1/\omega^2$ and μ, as follows:

$$(1/\tau_2^2)\cdot(1/\omega^2)(\mu-\mu_2-\mu_3)+(\mu-\mu_3)=0, \tag{3}$$

where $1/\tau_2^2$, μ_2, and μ_3 are constants. The graph of this equation is a hyperbola having a μ-intercept of μ_3, and asymptotes $\mu=\mu_2+\mu_3$ and $1/\omega^2=-\tau_2^2$. It was recognized that the downward concavity of the pertinent branch of this curve coincided well with the distribution

[1] A. Tobolsky and H. Eyring, J. Chem. Phys. **11**, 125 (1943).

[2] Gehman, Woodford, and Stambaugh, Ind. Eng. Chem. **33**, 1032 (1941).

[3] Dillon, Prettyman, and Hall, J. App. Phys. **15**, 309 (1944).

[4] W. J. Lyons and I. B. Prettyman, J. App. Phys. **19**, 473 (1948).

[5] W. J. Lyons, Textile Research J. **19**, 123 (1949).

of the experimental points in Fig. 1. Accordingly the three constants were evaluated to secure a good fit to these points. The resultant theoretical curve, for which $\tau_2=9.3$ sec., $\mu_2=12.4\times10^6$ poises, $\mu_3=1.13\times10^6$ poises, has been drawn in Fig. 1. The values of the constant coefficients in the equation set forth in Fig. 1 are those which are appropriate when $1/\omega_0$ is expressed in 10^{-3} sec.

DISCUSSION

The foregoing analysis has provided a value for the relaxation time τ_2 for secondary-bond processes. The value of 9.3 sec. is consistent with the conditions accepted by Tobolsky and Eyring in deriving Eq. (2), one of the conditions being that τ_2 is of the order of seconds.

If the familiar relation of Maxwellian relaxation theory, $\tau_i=\mu_i/E_i$, is considered as holding in intermolecular-bond processes, it is possible to estimate the value of the partial dynamic modulus E_2, for these processes. Thus

$$E_2=\mu_2/\tau_2 = (12.4\times10^6)/9.3=1.33\times10^6 \text{ dynes/cm}^2. \quad (4)$$

This value is practically negligible in comparison with the value of E which has been found for the Nylon sample: 8.0×10^{10} dynes/cm^2. It would appear then that in the type of longitudinal vibration to which the Nylon sample was subjected in the stretch-vibrometer, secondary bonds contribute very little to the over-all elasticity, and thus,

$$E\approx E_1. \quad (5)$$

The implication here that the primary bonds furnish practically all of the dynamic elasticity is at variance with the interpretation of Tobolsky and Eyring, who consider E_1 to be essentially the static modulus. Thus,

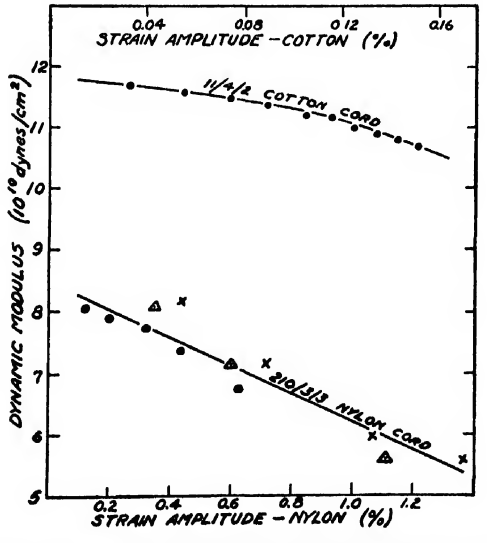


FIG. 2. Influence of strain-amplitude on the dynamic moduli of 11/4/2 cotton and 210/3/3 Nylon cord; three separate experiments on Nylon sample.

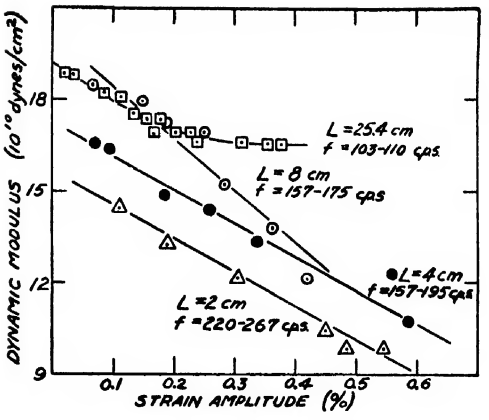


FIG. 3. Influence of strain-amplitude on the dynamic modulus of 1100/2 rayon cord; 2-kg load, at various free lengths of specimen, L.

on this view, E_1 and E_2 would be of the same order of magnitude, with E_1 being actually the smaller of the two, since this, quite generally, is what experiment indicates.

While the Tobolsky and Eyring analysis appears to be reasonable as applied to the rubber-like polymers, it is expected that in an oriented, linear fiber-forming polymer, such as the Nylon, being vibrated at a relatively small strain-amplitude, the basic dynamic, elastic processes could very well be restricted to the primary bonds. This behavior implies that the load is carried principally by the main valence chains, which view, for an oriented textile polymer such as high tenacity Nylon, is indeed tenable. In fact, Mark,[6] speaking of textile polymers, cellulose in particular, says "If any kind of van der Waals' forces were assumed to be responsible for the modulus of elasticity . . . the high values actually observed by various authors (would not be expected)." Secondary bonds, as represented by van der Waals' forces, are estimated to be smaller than primary valence forces by several orders of magnitude, in line with the relation found between E_1 and E_2 in the present analysis.

As the strains are carried to higher amplitudes it may be expected that more and more secondary bonds become involved, and the load is partially shifted over to them. This would result in an increase in the effective secondary-bond modulus E_2, and a decrease in the effective primary modulus E_1, but if replacement here is bond-for-bond, or even in the ratio of 100 for 1, the effect would be a decline in the over-all modulus E. This effect we have found in strain-amplitude studies on cotton, rayon and Nylon cords, as shown in Figs. 2 and 3.

Considered as an equation in μ and $1/\omega_0$, instead of $1/\omega_0^2$, Eq. (3) is a cubic. The graph of this cubic equation is plotted as a solid line in Fig. 4. By equating $d^2\mu/d(1/\omega_0)^2$ to zero the inflection point of the cubic is

[6] H. Mark, *Cellulose and Cellulose Derivatives* (Interscience Publishers, Inc., New York, 1943), edited by E. Ott, p. 995.

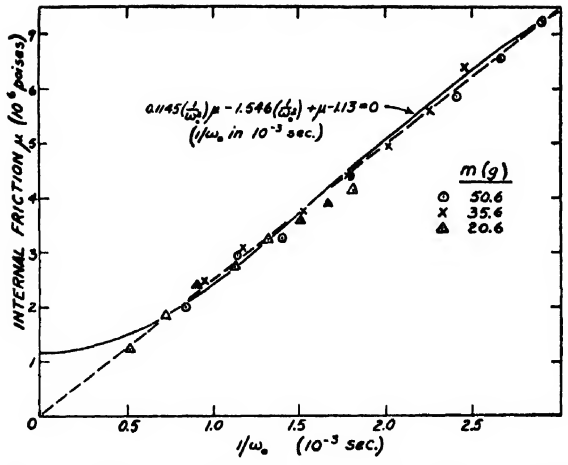


FIG. 4. Observed coefficient of internal friction of Nylon monofil as a function of $1/\omega_0$. Fitted to the data is the graph of Eq. (2) (solid line), as well as that of $\mu\omega_0 = \text{constant}$.

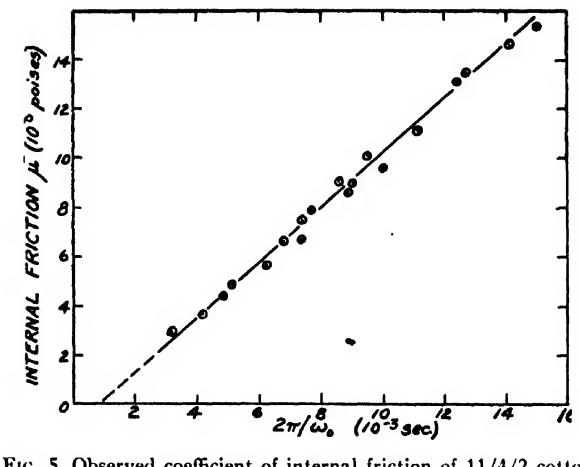


FIG. 5. Observed coefficient of internal friction of 11/4/2 cotton tirecord as a function of $2\pi/\omega_0$.

found to be at $1/\omega_0 = 1.7\times10^{-3}$ sec. The region of the inflection point of a cubic curve can be approximated by a straight line, and since the inflection point falls at about the middle of the array of experimental points in this study, the data themselves, as Fig. 4 shows, can be as well represented by a straight line as by the cubic. The equation of such a straight line is the hyperbolic relation

$$(\mu+\alpha)\omega_0 = \text{constant}, \tag{6}$$

with which, it has been demonstrated,[4,5] our data on a large number of textile samples are in apparent agreement. While Eq. (2) or (3) is not mathematically consistent with Eq. (6), the foregoing graphical analysis indicates why the data are in reasonably good agreement with both.

The straight line in Fig. 4, which is the most acceptable that could be drawn through the data points, has been extrapolated to the origin. The fact that a straight line passing through the origin fits the data, plotted on a $(1/\omega_0)\mu$-plane, demonstrates better than has been done heretofore that: (a) the relation between μ and ω_0 may be represented by a *true* hyperbola, and (b) the constant α is zero, and ω_0 has no additive constant. Furthermore, from Fig. 4 one is readily able to evaluate the constant in Eq. (6). Thus the data for the Nylon monofil, if they are regarded as representable by Eq. (6), lead to the relation

$$\mu\omega_0 = 2.47\times10^9 \text{ ergs/cm}^3. \tag{7}$$

An interesting sidelight emerges from the consideration of the constancy of the product $\mu\omega_0$, for, referring (5) to the equation

$$\mu = qF/s_m\omega_0, \tag{8}$$

it is seen that $\mu\omega_0 = \text{constant}$ implies that the displacement amplitude at resonance s_m is directly proportional to the amplitude of the applied force F, if the shape factor $q = L/2A$ remains constant. For a particular string, *strain*-amplitude is proportional to force-amplitude regardless of length L. Such a relationship appears entirely reasonable, though it is not part of the original dynamic theory. These considerations evidently place the product $\mu\omega_0$ in the role of a Young's modulus for the viscous component of the strain. This interpretation of $\mu\omega_0$ is brought out also (and without recourse to experiment) when the dynamics is expressed in complex notation.

COTTON CORD DATA

Following the latter treatment applied to the Nylon data, results on a cotton cord, taken from Fig. 6 of reference 4, have been plotted on a $(1/\omega_0)\mu$-plane, as shown in Fig. 5. Here again a straight line appears to represent the frequency dependence of μ. However, the line does not pass through the origin. The μ-intercept $(=-\alpha)$ is found to be -0.9. Evaluating the slope of the graph we get for the cotton sample,

$$(\mu+0.9)\omega_0 = 7.05\times10^9 \text{ ergs/cm}^3. \tag{9}$$

A natural question arises as to what theoretical interpretation is to be put on the finite value of α. In the light of the discussion on the Nylon data, it appears that no serious attempt to place a physical interpretation on α should be made. Evidently the vanishing of α in the Nylon case was quite fortuitous.

ACKNOWLEDGMENTS

The data upon which the present development rests are the results of the careful measurements and calculations of Mr. Irven B. Prettyman. The author has had the benefit of profitable discussions on the analytical aspects with Messrs. J. W. Liska and G. R. Thurman. The encouraging support of Firestone research management is gratefully acknowledged.

Equation of State for High Polymers

R. S. Spencer and G. D. Gilmore
The Dow Chemical Company, Midland, Michigan

In a former paper the authors suggested the following equation of state, $(P+\pi)(V-\omega)=RT$, where π is the internal pressure (or cohesive energy density) and ω is the volume at absolute zero. It was found that this equation was in good agreement with data on the thermal expansivity and isothermal compressibility of polystyrene.

The present paper reports isothermal compression measurements on polymethyl methacrylate, polyethylene, cellulose acetate-butyrate, and ethyl cellulose at temperatures within the fabrication range. The above equation of state is shown to apply equally as well to these thermoplastics as to polystyrene and values are given for the constants. The usefulness of this relationship is illustrated in a brief discussion of some phases of the injection molding process.

INTRODUCTION

In a former paper the authors suggested[1] the following modified van der Waals' equation of state,

$$(P+\pi)(V-\omega)=RT.$$

In this, P is the external pressure, π the internal pressure (or cohesive energy density), V is the volume, ω the volume at absolute zero, R the gas constant, and T the absolute temperature. It was found that this relationship was obeyed quite well by polystyrene, and served to correlate experimental observations of thermal expansion, isothermal compression, sound velocity, and cohesive energy density by swelling measurements.

This work was undertaken as part of a larger study of the fundamentals of the injection molding process, and most of our applications of this equation of state have been in that field. The desire to extend our findings to thermoplastics other than polystyrene has resulted in the present work, inasmuch as the equation of state is basic to further considerations in injection molding theory. This point will be illustrated briefly at the end of the discussion.

The pragmatic nature of our interest in this problem has limited considerably the scope of our investigation. No attempt was made to study the compressibility of these polymers below their apparent second-order transition temperatures, as the interest was centered on fabrication conditions. Preliminary attempts were made to observe time effects associated with pressure and temperature changes in the neighborhood of the apparent second-order transition,[2] but it was concluded that the apparatus used was unsuitable. It was noted, however, that polystyrene exhibits retarded isobaric volume changes, after a sudden change in pressure, as well as the retarded isothermal volume changes already reported.[2,3] It is hoped that this important phenomenon will be investigated further by other workers in the field.

We shall continue to use the hybrid units of reference 1 and express P and π in lb./sq. in., V and ω in cc/g, and T in degrees K.

EXPERIMENTAL METHODS

The basic apparatus used for isothermal compression measurements has been described elsewhere,[1] as follows: "It consists of a cylindrical specimen cavity, one square inch in cross-sectional area, with a close-fitting plunger for compressing the polymer. This plunger also contains a cavity which opens into a groove machined in the wall of the specimen cavity. This plunger cavity is filled with polymer, which is forced into the groove and acts as a seal for the sample in the main cavity. In operation, then, force is applied to the smaller central plunger, and the larger plunger directly contacting the sample *floats*."

Some slight changes have been made to facilitate operation. When it was desired to control temperature more closely than was possible with the electric heater the mold was placed inside a small insulated tank through which was pumped liquid from a thermostatted bath. The compression measurements were carried out with a Baldwin-Tate-Emery testing machine, which provided loads up to 20,000 lb. The compression strain was measured with a microformer-type (differential transformer) strain follower which operated the chart drive of a drum recorder. The pen movement was linked mechanically to the testing machine so as to record pressure. Thus, the recorder reproduced automatically the compression stress-strain curve. This arrangement effected a considerable speeding up of the measurements.

The polymethyl methacrylate used was a product of E. I. du Pont de Nemours and Company, designated as Lucite HM 140. The polyethylene was also a du Pont product, designated as P1000 PM 1. The cellulose acetate-butyrate was Tennessee Eastman's Tenite II, designated as 265-A. The ethyl cellulose formulation was a mixture of 15 percent plasticizer, 85 percent ethyl cellulose of degree of substitution 2.34. All of the samples were prepared by machining down cylinders to a firm press fit into the compressibility mold.

DISCUSSION

Broadly speaking, our measurements consist of parallel observations of pressure, P, and volume, V, at a fixed temperature, T. From such data we wish to deduce the probable values of the constants π, ω, and R.

[1] R. S. Spencer and G. D. Gilmore, J. App. Phys. **20**, 502–6 (1949).
[2] R. S. Spencer and R. F. Boyer, J. App. Phys. **17**, 398–404 (1946).
[3] Alfrey, Goldfinger, and Mark, J. App. Phys. **14**, 700 (1943).

TABLE I. Internal pressure of polymethyl methacrylate at 175°C.

$\bar{P}$	π
2500 p.s.i.	32,000 p.s.i.
3500	31,300
4500	31,700
5500	32,900
6500	30,100
7500	32,500
8500	27,200
9500	35,800
11,000	30,200
13,000	32,400
15,000	28,500
17,000	30,500
	Average = 31,300

TABLE II. Internal pressure of polyethylene at 178°C.

$\bar{P}$	π
1560 p.s.i.	46,500 p.s.i.
2500	47,700
3525	48,000
4605	47,400
5680	48,800
6600	46,400
7500	47,600
8540	47,600
9540	47,400
10,525	46,700
11,525	50,500
12,750	46,600
14,270	47,500
15,540	47,300
16,520	47,800
17,545	47,100
	Average = 47,600

This may be done, of course, by direct use of the equation of state in the form already given. However, there are experimental and theoretical advantages in considering only corresponding changes in pressure and volume, rather than absolute magnitudes. This procedure avoids the necessity of using the apparatus to measure thermal expansivity (for which it is not ideally suited), helps prevent the accumulation of errors, and is more sensitive in bringing out such things as the pressure-dependence of π.

In this differential method we eliminate ω, which can be calculated subsequently from a known value of the equilibrium volume at some temperature and pressure. The constants π and R remain to be determined from our data. If we knew the molecular weight, M, of an "interaction unit" of the polymer chain, we could calculate R, in our units, from our relationship $R=1206/M$. If the polymer is something like polystyrene or polymethyl methacrylate, having one bulky side-group per monomeric residue, it seems reasonable to assume that the interaction unit is just one residue. This appears to be true, within experimental error. Such polymers as polyethylene, with no side-groups, and cellulose derivatives, with several per residue, present much more ambiguous situations, however. It seems preferable, then, to begin by considering both π and R to be unknown constants. If the empirical value of R subsequently offers a suggestion as to the nature of the interaction unit, the data can be reworked accordingly.

The appropriate relationship to use in this approach is

$$[-T(\Delta P/\Delta V)]^{\frac{1}{2}}=[\pi/(R)^{\frac{1}{2}}]+[1/(R)^{\frac{1}{2}}]\cdot\bar{P}$$

where ΔV is the volume change resulting from a pressure change ΔP, and $\bar{P}$ is the average pressure over the interval ΔP. Probable values of the constants can be determined by drawing a straight line through a plot of $[-T(\Delta P/\Delta V)]^{\frac{1}{2}}$ *vs.* $\bar{P}$, or by a least-squares procedure, which was the method adopted.

If the monomeric residue is the interaction unit in polymethyl methacrylate, we should expect $R=12.05$. Applying a least-squares procedure to compression measurements at 175°C, we obtain $R=14.71$ and $\pi=35{,}300$ p.s.i. Assuming that the discrepancy in R is experimental error and taking $R=12.05$ as correct, the value of π can be computed over each pressure interval, as shown in Table I. The average value is $\pi=31{,}300$ p.s.i., the scatter is not excessive (there appears to have been an error in the volume of $P=9000$ p.s.i.) and there is no dependence of π on P over the range covered. The equilibrium volume-temperature coefficient would be $(\partial V/\partial T)_{P=0}=R/\pi=3.86\times10^{-4}$ cc/°C. The density at 25°C was 1.169 g/cc. Assuming this to be an equilibrium value, we calculate $\omega=0.740$ cc/g. If we calculate the density at the apparent second-order transition, using the data of Robinson, Ruggy, and Slantz,[4] and assume that to be a state of equilibrium, we obtain $\omega=0.734$. If we take the density at 30°C = 1.1818 g/cc, as given in reference 4, to be an equilibrium value, we get $\omega=0.728$. The last two values are probably preferable to the first one, only suitable equilibrium density measurements can resolve the question.

In the case of polyethylene at 178°C, the least-squares method gives $\pi=50{,}600$ p.s.i., $R=48.03$. This value of R is not far from 43.0, which corresponds to one ethylene residue as an interaction unit, so the latter value shall be taken as correct. Calculating for each pressure interval we obtain the results of Table II, which give as an average $\pi=47{,}600$ p.s.i. The equilibrium volume-temperature coefficient is $(\partial V/\partial T)_{P=0}=9.48\times10^{-4}$ cc/°C. The density at 25°C was found to be 0.9196 g/cc, from this $\omega=0.805$ cc/g. This is almost certainly not correct, however, due to the presence of considerable crystallinity at room temperature. Correcting the density to the melting point, using the data of Clash and Rynkiewicz,[5] we obtain $\omega=0.875$.

The least-squares values obtained with the data on ethyl cellulose at 196°C are $R=19.95$ and $\pi=34{,}800$ p.s.i. The value of R corresponds only to a fractional

[4] Robinson, Ruggy, and Slantz, J. App. Phys. **15**, 343–51 (1944).

[5] R. F. Clash, Jr., and L. M. Rynkiewicz, Ind. Eng. Chem. **36**, 279–82 (1944).

TABLE III. Internal pressure of ethyl cellulose at 196°C.

$\bar{P}$	π
1500 p.s.i.	34,700 p.s.i.
2500	34,900
3560	34,100
4560	35,600
5500	34,500
6500	35,300
7500	35,700
8500	33,700
9500	34,800
10,500	33,300
11,540	35,800
12,540	35,000
13,500	40,400
14,500	30,400
15,500	34,900
16,500	33,900
17,500	35,400
	Average = 34,800

TABLE IV. Internal pressure of cellulose acetate-butyrate at 181°C.

$\bar{P}$	π
1500 p.s.i.	41,800 p.s.i.
2500	40,000
3500	41,800
4500	42,500
5500	40,300
6500	41,800
7500	40,100
8500	41,000
9500	41,500
10,550	42,000
11,550	43,600
12,500	41,800
13,500	39,900
14,500	38,900
16,000	44,800
17,500	39,700
	Average = 41,300

part of a substituted glucose residue, as might be expected, therefore, we shall accept the least-squares value, having no better substitute. The scatter in the measurements is shown in Table III, where π is calculated for each pressure interval, using the above value of R. No trend of π with $\bar{P}$ may be noted. There appears to be an error in the volume observed at $P = 14{,}000$. The equilibrium volume-temperature coefficient is $(\partial V/\partial T)_{P=0} = 5.73 \times 10^{-4}$ cc/°C. The density at 25°C was observed to be 1.115 g/cc, giving $\omega = 0.725$. Correcting to the apparent second-order transition, using the data of Wiley,[6] gave $\omega = 0.720$.

In the case of cellulose acetate-butyrate at 181°C the least-squares values are $R = 22.18$ and $\pi = 41{,}300$ p.s.i. As with ethyl cellulose, the value of R corresponds to a fractional residue and will be accepted as given. In Table IV are shown the values of internal pressure for each pressure interval, calculated with the above value of R. The scatter and lack of dependence upon applied pressure are very much as has been noted in the preceding tables. The equilibrium volume-temperature coefficient is $(\partial V/\partial T)_{P=0} = 5.37 \times 10^{-4}$ cc/°C. The density at 25°C was 1.17 g/cc, giving $\omega = 0.695$. Correcting to equilibrium, using the data of Clash and Rynkiewicz,[5] gives $\omega = 0.688$.

The equation of state constants which have been determined to date are summarized in Table V, together with activation energies for viscous flow, taken from various sources.[4,7,8] The activation energy for flow seems to rise with increasing cohesive energy density, as might be expected, with the glaring exception of polyethylene. This exception is apparent rather than real, however. The units of π are energy per unit volume, while the units of activation energy are energy per mole. Taking the residue molecular weights into account, a fair correlation is found between cohesive energy per mole and activation energy per mole for all five polymers. In Eyring's picture of viscous flow[9] this would be interpreted as signifying that the number of residues per kinetic chain segment is roughly a constant for the different polymers. We estimate that it is in the general neighborhood of 15 residues per segment.

The order in which the polymers arrange themselves on the basis of cohesive energy density is not too surprising, again with the exception of polyethylene. Considering only chemical structure, one would be tempted to guess, *a priori*, that polyethylene would have a lower cohesive energy density than the other polymers, and just the reverse is true, of course. However, the polyethylene molecule exhibits a quality which the others do not, and that is great regularity of structure and the ability to form close-packed arrangements. It is just this factor which permits polyethylene to crystallize in spite of its non-polar nature. Thus, polyethylene contains many more monomeric residues per cc than the other polymers. If we calculate the cohesive energy per residue, or per interaction unit, the apparent discrepancy disappears.

In discussing a few examples of the important role which the equation of state plays in injection molding theory it will be necessary first to present a somewhat simplified picture of what occurs in an injection molding cycle.[10] In the more usual type of molding press the polymer is fed in at room temperature and is heated to molding temperature in the machine. In a subsequent cycle this hot polymer is forced through the channels of the press and mold, eventually filling the mold cavity. It is customary to maintain pressure on the ram of the press for some time after the mold is filled. During this interval several things are going on. The pressure in the mold builds up fairly rapidly to a

[6] F. E. Wiley, Ind. Eng. Chem. 34, 1052-6 (1942).
[7] G. J. Dienes, J. Colloid Sci. 2, 131-61 (1947).
[8] R. S. Spencer, J. Polymer Sci. (to be published).
[9] Glasstone, Laidler, and Eyring, *The Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941), p. 500.
[10] G. D. Gilmore and R. S. Spencer, Mod. Plastics (to be published).

TABLE V. Equation of state constants for high polymers.

Polymer	R	ω	π	Viscosity activation energy
Polystyrene	11.6	0.822 cc/g	27,000 p.s.i.	23 kcal./mole
Polymethyl methacrylate	12.05	0.734[a]	31,300	30
Ethyl cellulose	19.95	0.720[a]	34,800	40
Cellulose acetate-butyrate	22.18	0.688[a]	41,300	78
Polyethylene	43.0	0.875[a]	47,600	11

[a] Values uncertain owing to lack of reliable values for equilibrium density.

relatively stable value, which is usually appreciably lower than the ram pressure. Inasmuch as the mold is colder than the polymer, the polymer is being cooled from the time it enters the mold. The mass of the polymer in the mold increases slowly during this time. At the end of this interval the ram is returned to its starting position. The pressure difference between the mold cavity and the channels in the press has now reversed direction and there is certain amount of polymer discharged from the mold cavity. This discharge lowers the pressure in the mold and the discharge rate falls off until there is sufficient cooling at the constricted entrance to the mold cavity (usually called the "gate") to "solidify" the polymer and seal off the mold. Following the discharge period the polymer in the mold cools under conditions of constant mass until the mold is opened and the piece removed.

The cooling characteristics of the mold give us a relationship between polymer temperature and time. The equation of state then gives us the relationship between pressure, mass of polymer in the mold, and temperature. Thus, consideration of any question which involves calculating the pressure or the mass of polymer in the mold must rest on a knowledge of the equation of state. For example, under certain conditions volume defects, such as bubbles or sink marks, appear in the piece. These defects can be rendered unobjectionable by being sure that the mass of polymer sealed in the mold exceeds some minimum value under given sealing conditions. This condition can be set up in terms of controllable quantities by use of the equation of state. As another example, we note that behavior during release of the piece from the mold is determined primarily by the pressure remaining in the mold at the time the mold is opened. The force required to open the mold depends directly upon the residual pressure, as might be expected. If the pressure is too high it will not be possible to open the mold. At lower pressures the piece will stick in the female half of the mold and/or will be objectionably scored in removing from the mold. Finally, if the residual pressure is kept below a certain maximum, the piece releases reasonably easily. If, however, the pressure is allowed to become "negative," there is a tendency to stick on the male half of the mold, and even to crack. The application of the equation of state in considering such questions as the above should be obvious.

A more detailed discussion of these, and other, injection molding problems may be found in reference 10 and in publications to follow.

SUMMARY

The constants in a simplified van der Waal's equation of state have now been determined for the following polymers: polystyrene, polymethyl methacrylate, ethyl cellulose, cellulose acetate-butyrate, and polyethylene. Of these, polystyrene has the lowest internal pressure (or cohesive energy density) and polyethylene has the highest. The unexpectedly high internal pressure exhibited by polyethylene has been associated with the great regularity of the polyethylene molecule and its ability to form closely packed arrangements.

The application of the equation of state in the injection molding field has been illustrated by a very brief discussion of the roles played by pressure, temperature and polymer density in such phenomena as mold release and the formation of bubbles and other defects.

Behavior of Granulated Polymers under Pressure

R. S. Spencer, G. D. Gilmore, and R. M. Wiley
The Dow Chemical Company, Midland, Michigan

In such fabricating processes as extrusion and injection molding it is frequently necessary to force a granular polymer through a channel. Frictional forces between polymer and channel walls lead to a drop in pressure through the granular material. Simultaneously, the applied pressure causes a compaction of the granules.

Assuming that the granular mass has certain fluid-like properties, a simple theoretical expression for pressure transmission is derived. Comparison with data on saran powders and granular polystyrenes shows good agreement between theory and experiment. The effect of lubrication is shown also.

The compaction of granular polystyrene follows the same pattern as that observed in the earth's crust. The porosity decreases exponentially with increasing pressure. The "compressibility coefficient" for the porosity of granular polystyrene is comparable with that given for clays and shales.

INTRODUCTION

THE available literature relating to the behavior of granular solids under pressure is rather scanty. Inasmuch as the handling of some such materials is of considerable technological importance, the reasons for this deficiency are not immediately apparent. It can only be surmised that either an appreciable amount of unreported work has been done or else the practical importance of the problem has not been emphasized sufficiently to those who are in a position to study it adequately.

Most thermoplastics are marketed in granular form. In subsequent fabrication operations, such as extrusion or injection molding, it is usually necessary to force the granular polymer through some sort of channel. In given circumstances, then, one wishes to know what force it is necessary to apply in order to move the granular polymer against a specified resisting force, or, conversely, how much of the applied pressure is transmitted through the granular material. It is probably safe to say that lack of knowledge on this point is not least among the factors which have served to maintain polymer fabrication techniques in the realm of technological arts rather than applied sciences.

The problem to be considered in the present paper is that of the relationship between applied force and resisting force at either end of a plug of granular polymer which is just on the point of motion through a channel of constant, but arbitrary, cross section. That is to say, the conditions at the wall of the channel are those of static friction. The solution of this problem necessitates certain assumptions, which are set up largely by analogy with fluid behavior. This restricts application to those cases in which the particle size is considerably smaller than the dimensions of the channel. Two simple arrangements will be described by which the problem may be studied experimentally, each having certain advantages. Results obtained with saran powders and polystyrene granules will be presented and interpreted with the aid of the simple theory.

At relatively higher pressures, considerable compaction of the granular material may be effected, which is essentially irreversible as long as the plug is confined within the channel. This is taken account of in the assumptions underlying the theory and, experimentally, by always proceeding in the direction of increasing applied pressure. This compaction is of some interest in itself; therefore, a few data are presented and discussed to illustrate the pattern of behavior.

It is doubtless redundant to point out that the treatment throughout is simplified and the investigation undetailed. Our interest in this problem arose because of its place in a larger program of study of polymer fabrication processes. Many points, theoretical and experimental, await clarification. It may be hoped that this will not be too long forthcoming.

EXPERIMENTAL METHODS

The first arrangement, being that used to study saran powders, is shown in Fig. 1. The confining channel was a glass tube of circular cross section. The polymer sample rested upon a rod extending below the tube, and was confined from above by a weight representing the resisting force. Both weight and rod were smaller in diameter than the tube, flaring out to the tube diameter at the end contacting the polymer. This was to minimize binding which would otherwise be occasioned by the trapping of powder granules between the rod and tube. One arrangement utilized a tube of 1.1 cm internal diameter with weights of 17 and 35 g; another, a tube of 2.15 cm internal diameter with weights of 38 and 78 g.

The procedure followed was to assemble the apparatus, using a given amount of saran powder, and set it upright on the platform of a spring scale. Downward force was then exerted upon the glass tube, and the reading of the scale noted at the instant the polymer first began to slip. This was then repeated with different amounts of powder, resisting loads, and tube diameters.

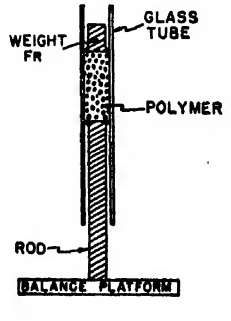


FIG. 1. Experimental arrangement used in the study of saran powders.

TABLE I. Coefficient of friction of saran powders.

Powder No.	D (cm)	μ
1	1.1	0.215
	2.15	0.239
2	1.1	0.144
	2.15	0.144
3	1.1	0.113
	2.15	0.167

It was found that all of the glass tubes examined were not perfectly cylindrical but had varying degrees of taper. It was necessary, then, to select those with the least taper, and to exercise the further precaution of always moving the polymer along the tube in the direction of increasing internal diameter, to avoid jamming. Relatively reproducible results could then be obtained with some care, although experimental errors with this apparatus must be considered to be rather large.

The apparatus used for the study of granular polystyrenes is shown schematically in Fig. 2. It consisted of a heavy-walled steel cylinder of 1.953 in. internal diameter, with two closely fitting pistons. SR-4 strain gauges were mounted on the lower piston and the unit calibrated in terms of force applied to the piston. It was found that any film on the surface of the cylinder wall adversely affected the reproducibility of measurement, therefore, the unit was carefully cleaned with carbon tetrachloride before use. It was then assembled, with the sample in place, and set between the platens of a 30-ton press. Force was applied through a hand-operated hydraulic jack, and the corresponding transmitted force determined from the strain gauge readings.

The saran powders used were commercial materials, denoted as *1*, *2*, and *3*, in order of increasing amount of lubricant in the formulation. The polystyrenes were also commercial materials, very similar to each other. The sarans were in the form of very finely divided powders, whereas the polystyrenes were much coarser, perhaps averaging around 0.15 in. particle diameter. One exception to this was polystyrene No. 2, which was that component of No. 1 which would pass through a U. S. No. 8 screen, and averaged considerably less than that in size, perhaps around 0.04 in.

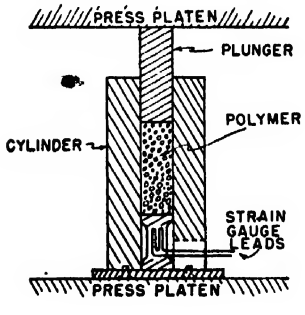


FIG. 2. Experimental arrangement used in the study of pressure transmission through granular polystyrenes.

DISCUSSION

The pressure-transmission problem to be solved has already been stated in the Introduction; it may be made a little clearer, perhaps, by referring to Fig. 3. A plug of granular polymer is confined within a channel of constant cross section. The applied force at one end of the plug is in opposition to the resisting force at the other end. What is desired is the relationship between the forces when the plug is just on the point of moving in the direction of the applied force.

It has been mentioned that increasing the applied force causes compaction of the granules. That is to say, the length of the plug decreases. The state of compaction may be defined by the porosity, or the initial and currently observed lengths of the plug, or by the initial length and the current pair of forces. In general, we must consider that the porosity will vary along the length of the plug; however, we shall be considering in our solution only the equilibrium situation in which the porosity at any point is not changing with time.

We shall define now the following quantities:

x = distance from the forward end of the plug to any plane normal to the axis of the plug.
μ = coefficient of friction of particle on wall.
n = number of particles per unit area contacting the wall at x.
n_0 = number of particles per unit area contacting the piston surfaces at the ends, or in contact over a cross section.
f = contact force between particles at x.
F = total force normal to the cross section at x.
$F_A(P_A)$ = applied force (pressure) necessary to just move the plug.
$F_R(P_R)$ = resisting force (pressure).
L = length of the granular plug in any given state of compaction.
L_0 = original length of the plug when the forces were very small.
C = perimeter of channel cross section.
A = area of channel cross section.
D = diameter of circular cross section
a, b = sides of rectangular cross section.

The principal assumptions to be made use of are the following:

(1) In a given state of compaction there is no tendency for particles to move with respect to each other, hence the only frictional forces to be considered arise from particle-wall contacts. (2) Contact forces are the same in all directions at a point, and are normal to the walls and ends. (3) The quantity n_0 is constant, independent of F and x. (4) The total number of wall contacts is constant, although, of course, n depends upon F (or upon x in a given state of compaction); further, that $n = n_0$ over the entire wall originally, when the forces are very small.

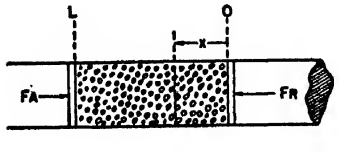


FIG. 3. Diagram defining quantities in the pressure transmission problem. A = area; C = perimeter.

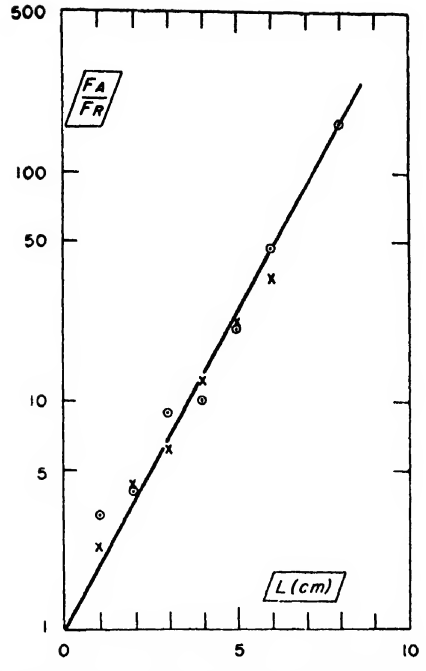


FIG. 4. Dependence of force ratio on length of sample of saran No. 2 in tube of diameter 1.1 cm. Circles represent $F_R=17$ g and crosses represent $F_R=35$ g.

Before proceeding, let us look over the assumptions quickly. There might be some temptation to replace (1) by the more usual assumption that the particles are hard and smooth, i.e., that there is no friction between particles. This is not true, however, as evidenced by the irreversibility of the compaction; thus we prefer the less general assumption (1). The second assumption is arrived at by considering the contact force to be analogous with pressure in a fluid. Actually, assumption (2) is more general than would be required, but it is felt that the analogy is clearer in the form as stated. The origin of assumptions (3) and (4) is more obscure and their plausibility less immediately discernible. It may help to picture the plug as a cylinder of some compressible material, on the surface of which is drawn an evenly spaced network representing the contact points. If this cylinder were compressed axially, while confined radially, the network spacing on the ends would remain unchanged, while the density of network on the sides would increase with the compression strain. Carrying this picture over to our granular material, we need only remember that in calculating the frictional force we are not concerned with the area of contact, but only with the contact force and the number of contacts.

With these preliminaries the solution is now quite straight-forward. Consider an element between x and $x+dx$. The frictional force at the wall is μf per particle, giving a total frictional force of $\mu fnCdx$. The change in F in going from x to $x+dx$ is $dF=An_0df$. By assumption (1) we can equate these two expressions and rearrange to give

$$dF/F=df/f=(\mu Cn/An_0)dx.$$

Integrating from $x=0$ to $x=L$, we obtain

$$\ln F_A/F_R=\frac{\mu C}{An_0}\int_0^L ndx.$$

To evaluate the integral on the right-hand side we utilize assumption (4) and equate the expressions for the number of wall contacts in the uncompacted state L_0 to the number of wall contacts in the compacted state. Thus,

$$n_0CL_0=\int_0^L nCdx$$

or

$$\int_0^L ndx=n_0L_0.$$

Substituting this in our previous equation, we finally arrive at

$$F_A/F_R=P_A/P_R=\exp(\mu CL_0/A).$$

We note that the exponent is merely the coefficient of friction times the ratio of the wall area (uncompacted) to the cross-sectional area. If the cross section is a circle of diameter D, we have

$$F_A/F_R=\exp(4\mu L_0/D).$$

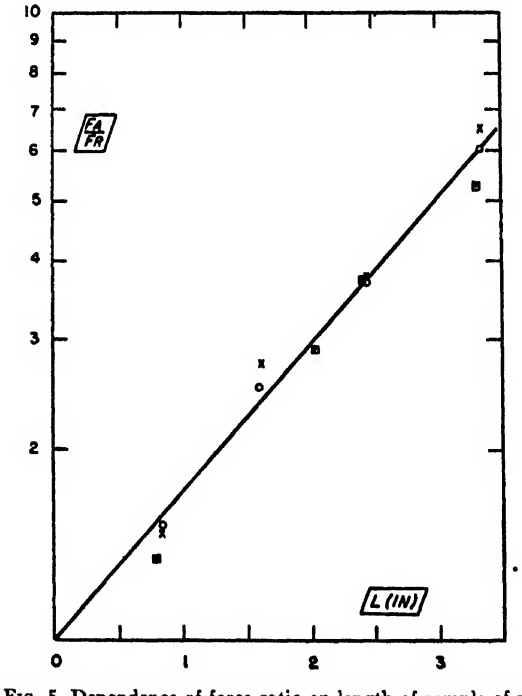


FIG. 5. Dependence of force ratio on length of sample of polystyrene in a steel cylinder of diameter 1.955 in. Circles represent polymer No. 1, crosses, No. 2, and squares, No. 3.

TABLE II. Pressure transmission through granular polystyrene. Force ratio (F_A/F_R).

Polymer No.	L_0(in.)	F_A=7600 lb.	14,700	22,000	29,500	37,000	44,000	51,000	59,000	Average
1	0.834	1.582	1.530	1.469	1.538	1.457	—	—	—	1.515
	1.594	2.71	2.37	2.44	2.57	2.64	2.51	2.45	2.46	2.52
	2.442	3.80	3.86	3.67	3.69	3.63	3.73	3.64	3.74	3.72
	3.354	6.33	5.65	5.79	6.15	6.17	6.29	6.22	5.57	6.02
2	0.829	1.520	1.362	1.448	1.490	1.516	—	—	—	1.467
	1.611	2.92	2.94	2.75	2.78	2.72	2.68	2.58	2.61	2.75
	2.447	3.80	3.34	3.67	3.69	3.85	3.93	3.98	3.74	3.75
	3.349	9.50	7.35	6.46	6.56	6.38	6.47	6.37	6.55	(6.47)
3	0.794	1.226	1.362	1.266	1.461	1.423	—	—	—	1.348
	2.042	3.17	2.83	2.75	2.74	2.72	2.78	2.96	2.98	2.87
	2.417	4.22	3.68	4.07	3.28	3.43	3.60	3.64	3.83	3.72
	3.314	3.46	3.68	3.66	3.99	4.63	4.89	5.31	5.27	(5.27)

TABLE III. Coefficient of friction of polystyrene on tool steel (Roche).

Temperature (°C)	μ
25	0.248
55	0.183
65	0.164
75	0.118
85	0.099

TABLE IV. Effect of surface lubrication of polystyrene granules.

PPM of lubricant	F_A/F_R
0	4.38
150	3.08
300	2.66
500	2.23
750	1.82

If it is a rectangle of sides a and b, we have

$$F_A/F_R=\exp[2\mu(a+b)L_0/ab].$$

If the rectangle approaches a slit, so that $a\gg b$, our expression approaches

$$F_A/F_R=\exp(2\mu L_0/b).$$

With the above results of our semicontinuous-semiparticulate approach to the problem, let us look at the experimental observations. First, we shall consider the data obtained on saran powders with the arrangement of Fig. 1. In Fig. 4 $\log F_A/F_R$ is plotted against L_0, for constant tube diameter and two resisting forces, and it may be noted that a fairly good straight line is obtained, as predicted by the theory. In Table I the data on the saran powders are summarized in the form of values of the coefficient of friction, calculated from the slopes of lines like that of Fig. 4. The constancy of F_A/F_R for different values of F_R is evident in Fig. 4 and the point is not elaborated further in Table I. It must be concluded that the saran data bear out the predicted dependence of F_A upon F_R, L_0 and D. The coefficient of friction of saran on glass is not sufficiently well known to provide a further test, but at least the values in Table I seem reasonable and decrease with increasing lubricant.

In the work with polystyrene, using the apparatus of Fig. 2, data were obtained in somewhat different form. Here it was convenient to use a number of different loadings and thus check in greater detail the predicted constancy of F_A/F_R as the loading increased and compaction proceeded. Quite high loads were possible, so that compaction down to as little as 60 percent of the original volume was achieved, yet F_A/F_R remained relatively constant, as shown in Table II. The effect of original length is shown in Fig. 5, where the average F_A/F_R from Table II is plotted logarithmically against L_0. Not only does a fair straight line result, but it may be seen that all three polystyrenes fall along the same line, which corresponds to a coefficient of friction of $\mu=0.262$. In Table III, we see experimental values for

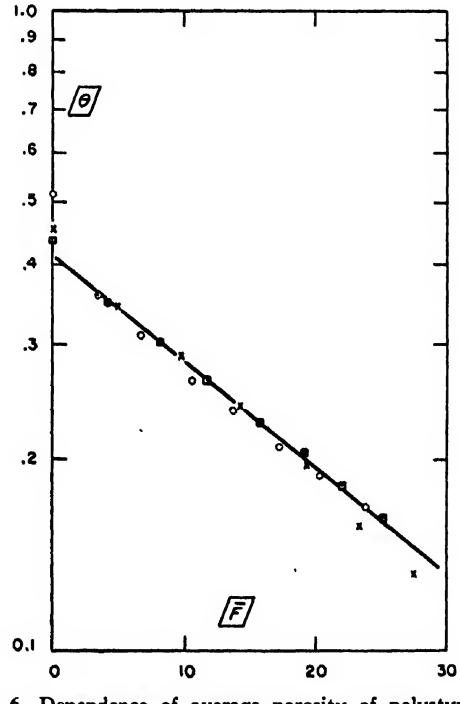


FIG. 6. Dependence of average porosity of polystyrene on average force. Circles represent 140 g of polymer No. 1, 90 g of No. 3, and squares, 140 g of No. 3.

the coefficient of friction of polystyrene on tool steel, as obtained by Roche.[1] Our value of 0.262 compares well with his value of 0.248 at 25°C, providing further evidence that the theoretical expression predicts observed behavior sufficiently well for practical applications.

It has been mentioned that it was necessary to clean the cylinder wall carefully in order to get reproducible results in the work with polystyrene. To see how great the effect of a lubricating film could be, one run was made with polystyrene No. 1 in which the wall was coated generously with Dow Corning No. 200 fluid. This reduced the coefficient of friction to 0.080. The effect of external lubrication on the polystyrene granules is similar. Different amounts of a lubricant in powder form were mixed thoroughly with polystyrene and the pressure transmission through a fixed size of sample determined. The results are shown in Table IV, where it is seen that increasing amount of lubricant reduces the coefficient of friction and thus reduces the ratio F_A/F_R.

Thus far little has been said about the compaction effect, except as incidental to the pressure-transmission problem. For some purposes a more quantitative formulation of the observed behavior may be desirable. The work of Athy[2] on the compaction of the earth's crust with depth suggests that a granular material may behave as if its free volume were characterized by a compressibility coefficient. That is, the porosity, θ, would follow the relationship $\theta=\theta_0 e^{-\beta P}$ where θ_0 is the initial porosity when the pressure, P, is zero, and β is a "compressibility coefficient." (Actually, we shall continue to talk about force, rather than pressure.) In our polystyrene experiments the force varies considerably over the length of the sample, and the porosity will vary accordingly. To correspond to the average porosity calculated from the length of the polymer plug we must calculate an appropriate average force, and this requires that we know the distribution of force along the plug. Assuming this to be of the form $F=F_R e^{kz}$ where k may be a function of L, L_0, etc., we find that the volume-average force is

$$\bar{F}=(F_A-F_R)/\ln(F_A/F_R).$$

Table V summarizes the results on compaction, and in Fig. 6 is plotted $\log\theta$ *vs.* $\bar{F}$. Except for an initial discrepancy, a straight line results which corresponds to a "compressibility coefficient" of $\beta=1.1\times10^{-4}$ in.²/lb. Using Athy's figures for compaction as a function of depth, and taking the mean density of the earth as 5.522 g/cc, we arrive at a value for the earth's crust of $\beta=1.25\times10^{-4}$ in.²/lb. One might surmise from this that the exact behavior in compaction of an aggregate of irregular particles is not very dependent on the nature of the particles.

TABLE V. Compaction of granular polystyrene under pressure.

Sample No.	Weight	$F_A\times10^{-3}$	$F_R\times10^{-3}$	$\bar{F}\times10^{-3}$	θ
1	140 g	0 lb.	0 lb.	0 lb.	0.511
		7.6	1.2	3.47	0.357
		14.7	2.3	6.69	0.309
		22.0	3.9	10.56	0.264
		29.5	4.9	13.70	0.235
		37.0	6.1	17.14	0.207
		44.0	7.2	20.3	0.186
		51.0	8.6	23.8	0.166
2	90	0	0	0	0.453
		7.6	3.0	4.96	0.345
		14.7	6.0	9.71	0.286
		22.0	8.4	14.13	0.240
		29.5	11.6	19.22	0.194
		37.0	13.6	23.4	0.155
		44.0	15.6	27.4	0.130
3	140	0	0	0	0.437
		7.6	1.8	4.03	0.350
		14.7	3.9	8.14	0.303
		22.0	5.3	11.74	0.265
		29.5	7.0	15.64	0.228
		37.0	8.2	19.10	0.203
		44.0	9.0	22.04	0.181
		51.0	10.0	25.15	0.161

SUMMARY

The problem considered was that of the relationship between the forces at the ends of a volume of granular polymer confined in a channel of constant cross section, when the granular mass is just on the point of motion. A solution of this problem was obtained on the basis of assumptions which were drawn, in part, from analogies with fluid behavior. Experiments were carried out with saran powders and granular polystyrenes which served to confirm the details of the theoretical expression within experimental error. The addition of lubricants to either the original polymer or the channel wall, was found to lower the coefficient of friction, as would be expected, and alter the forces correspondingly. As the forces at the ends were increased the length of the sample was observed to decrease. This compaction proceeds as if the average porosity of the sample were characterized by a compressibility coefficient, if an appropriate average force is taken. The magnitude of this coefficient for polystyrene is close to that observed in the compaction of the earth's crust, suggesting that it is the particulate nature of the material which is important, rather than the details of the individual particles.

[1] A. F. Roche, The Dow Chemical Company, Midland, Michigan (unpublished observations).
[2] L. F. Athy, Bull. Am. Assoc. Petroleum Geol. 14, 1-24 (1930).

Application of the Membrane Analogy to the Solution of Heat-Conduction Problems

L. H. Wilson* and A. J. Miles
Mechanical Engineering Department, Missouri School of Mines, Rolla, Missouri
(Received August 25, 1948)

This paper establishes the feasibility of applying the membrane analogy to the solution of two-dimensional steady-state heat-conduction problems. The membrane analogy is established, and the construction and the use of the apparatus for measuring the film are described. Inexpensive apparatus and simplicity of execution are the predominating features of this type solution.

NOMENCLATURE

t—temperature
x—space coordinate
y—space coordinate
z—vertical deflection of membrane
q—uniform pressure against membrane
S—surface tension of the membrane.

INTRODUCTION

In the solution of heat-conduction problems, the essential consideration is a determination of the temperature field. When irregular geometric configurations are encountered this problem becomes quite difficult and analytical approaches are laborious and at times impossible. As a result of this difficulty, recourse is made, when possible, to graphical or experimental methods, or to methods of analogy. It is the purpose of this paper to call to the attention of the reader a method of determining temperature fields in homogeneous heat-conducting members which has been heretofore neglected.

Consider a homogeneous membrane such as a soap film of any shape supported at the edges, subjected to a uniform tension at the edge, and subjected to a uniform lateral pressure. It may be shown that the mathematical expression for the membrane is as follows:

$$q/S=(\partial^2 z/\partial x^2)+(\partial^2 z/\partial y^2), \qquad (I)$$

where q is the pressure per unit area of the membrane, S is the uniform tension per unit length of its boundary, and z is the vertical deflection of the membrane. If there is no pressure exerted on either side of the membrane ($q=0$), the equation becomes

$$(\partial^2 z/\partial x^2)+(\partial^2 z/\partial y^2)=0. \qquad (II)$$

It was pointed out by Prandtl in 1903 that the equation which represents the torsion of an elastic bar of any cross section is identical with Eq. (I) as given above. In 1917, Taylor and Griffith (3)† first used the membrane analogy for the solution of torsion problems. They found that the maximum error encountered under specified conditions was 2 percent. The method was applied to determine the torsional stiffness and strength of cylindrical bars (4), flexure of beams (5), and the torsion and flexure of hollow shafts (6).

Work of a similar nature to that of Taylor and Griffith was performed by Trayer and March (14) with essentially the same equipment and comparable results.

In 1947, Neubauer and Boston (11) applied the membrane analogy to determine the torsional stress analysis of twist-drill sections.

The membrane analogy was used by Miles and Stephenson (9) in 1937 for the solution of a quite different problem. They recognized that the differential equation representing the pressure distribution about an oil or gas well is identical with Laplace's equation. They were able to determine the pressure distribution in oil and gas reservoirs of irregular shape when the reservoir was drained by a number of wells.

It was pointed out by Christopherson and Southwell (1) that every plane-potential problem has a membrane analog and that Prandtl's analog is merely a particular example which was concerned only with Laplace's plane-harmonic equation.

Consider the general equation for steady-state heat conduction in two dimensions.

$$(\partial^2 t/\partial x^2)+(\partial^2 t/\partial y^2)=0. \qquad (III)$$

From the foregoing discussion it may be seen that the membrane analogy is also applicable to the solution of two-dimensional steady-state heat-conduction problems. By comparing Eq. (II) with Eq. (III) it may be seen that the equations are identical when the temperature, t, in Eq. (III) is replaced by the vertical deflection of the membrane, z, in Eq. (II). Thus it is feasible that the deflection of the membrane represents, to some scale, the temperature distribution throughout the member under consideration, and if the boundary conditions are known (which is generally the case), a model of the heat-conducting member may be constructed accordingly.

* Present address: School of Mechanical Engineering, Purdue University, West Lafayette, Indiana.
† Numbers in parentheses refer to bibliography at the end of the paper.

DISCUSSION OF EXPERIMENTAL PROCEDURE

The apparatus employed was patterned after that used by Miles and Stephenson (9) except that the component parts were cast from aluminum. This change was made in order to reduce contamination of the film caused by oxidation of the metal. The instrument is shown in Fig. 1. The principal parts consisted of an aluminum shell constructed in two sections between which a plate may be inserted; an aluminum model of the member under consideration; a glass plate with a micrometer head projecting through the center of the glass, the plate being large enough to permit the micrometer gauge to be moved to any position in the shell without uncovering any portion of it; and a hinged target (which could always be brought down into the same position relative to the film under investigation), to which was clamped a paper disk on which points on contour lines were plotted. The base of the apparatus was equipped with three leveling screws to facilitate leveling or tilting to any desired angle. The micrometer head was ground to a point on the lower end so that it could be used as a probe to establish contact with the film. The upper end of the micrometer head was fitted with another needle point along the same axis as the lower point in order to mark the position of the probe on the paper target.

The soap solution used was prepared by mixing 2 grams of sodium oleate and 30 cc of glycerin in one liter of distilled water. Care should be taken to keep the soap solution free from impurities as these shorten the life of the membrane.

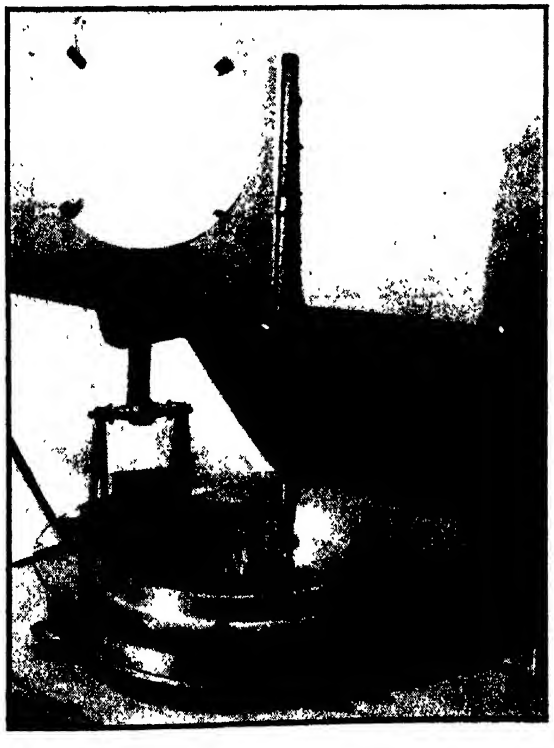


FIG. 1. Membrane apparatus with model of annulus-type heat-conducting member in place.

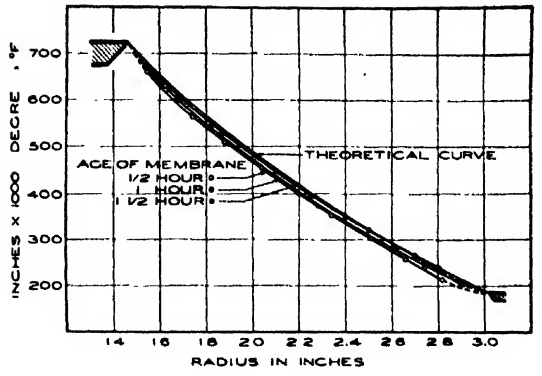


FIG. 2. Temperature distribution through thick-wall cylinder. $r_1 = 1.46$ in., $r_2 = 3.02$ in., $t_1 = 937°F$, $t_2 = 186°F$, slope of membrane = 48.2 percent, plate inclination 0.0°.

The models were constructed to some convenient scale geometrically similar to the heat-conducting member to be studied and in such a fashion that they provided a support for the membrane at the boundaries. These supporting edges were at different elevations such that they reproduced to some scale the temperature profile at the boundaries of the member under investigation. The membrane was then stretched across these boundaries, the elevation of the film at any particular point thereby representing the temperature at that point in the prototype of the member. Figure 1 shows a model of an annular-type heat-conducting member with isothermal boundaries, the inner boundary being at a higher temperature than the outer boundary. The model was made of aluminum to reduce oxidation, and the edges to which the membrane was to be attached were beveled in such a manner that the tangent to the membrane at its point of contact with the edge bisected the angle of bevel. This was necessary in order to reduce the possibility of the membrane's clinging to surfaces other than the beveled edge. When the model was set into the shell, the elevation of all edges representing isothermal conditions was checked with the micrometer head. If they were not level they were adjusted by using small jacks or weights to bring them into place.

With the model in place, a quantity of the soap solution was poured into the shell, care being taken to wet all surfaces of the model. With the aid of a Celluloid strip, a film was drawn over the model. All parts that come into contact with the soap solution were thoroughly cleaned and rinsed with distilled water beforehand. When the film had been drawn over the model, the glass plate was immediately placed over the equipment, since the carbon dioxide from the breath of persons present and also the dust in the air cause damage to the film. Arrangements were made during the manufacture of the model to provide ventilation holes. This was done in order that the pressure changes caused by temperature variation within the closed shell could

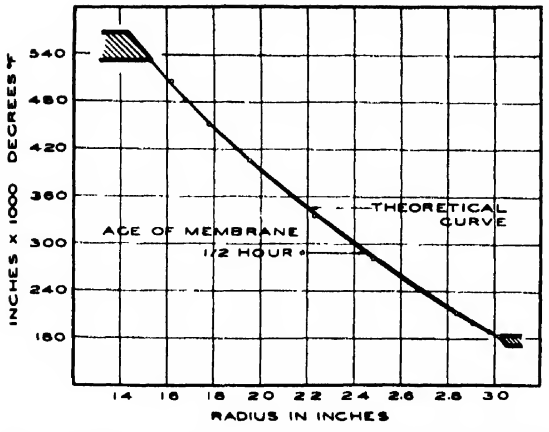


FIG. 3. Temperature distribution through thick-wall cylinder. $r_1=1.53$ in., $r_2=3.02$ in., $t_1=532.5°F$, $t_2=184°F$, slope of membrane=23.3 percent, plate inclination for drainage 3°, measurements taken on high side.

be equalized on both sides of the membrane. Small bubbles and excess liquid were removed from the membrane by using a saturated pipe cleaner or a soda straw. The membrane should have all such bubbles removed so that the surface is perfectly smooth.

In order to establish contour lines (isotherms) the micrometer was set at some specific height and the glass plate moved around until the lower tip of the micrometer touched the membrane at the desired point. The points of contact were found very accurately by having a light placed in such a position that the reflection of the lower point of the micrometer could be seen in the soap film near the place where the probe was expected to come into contact with the film. A distortion of the image showed that contact had occurred. This position was quite definite, and could be measured within plus or minus 0.001 inch. When contact had been made, the paper disk, which was clamped to the metal target, was brought down and touched lightly to the micrometer head. Contour lines were drawn by connecting all points on the isotherm. Extreme care must be exercised in keeping the lower point of the probe moistened with clean soap solution so that the film will not be destroyed when measurements are made.

Best results will be obtained if the apparatus is housed within a vibration-free building in which the temperature is kept nearly constant.

In order to compare experimental results with theoretical values, a problem for which the solution is known was first investigated.

Consider a hollow, thick-walled cylinder of inside radius, r_1, and outside radius r_2, and an inside temperature of t_1 with a lower outside temperature of t_2. It may be shown that the expression for the temperature, t, at any point, r, in the cylinder is as follows:

$$t=(t_2-t_1)\frac{\ln(r/r_1)}{\ln(r_2/r_1)}+t_1. \qquad \text{(IV)}$$

A model was constructed with an inside radius (r_1) of 1.45 inches and an outside radius (r_2) of 3.01 inches (see Fig. 1). The elevation of the outside shell was fixed at an elevation such that the reading on the micrometer head was 0.185 inch. The temperature at this point was assumed to be 185°F. The inner shell (r_1) was constructed such that the elevation could be varied; that is, the temperature drop through the cylinder could vary. It should be remembered that the difference in elevation of the two plates may represent any temperature difference desired, but in this experiment various elevations of the inside plate were used in order to determine the most appropriate slope of the membrane. Figure 2 shows typical results obtained when observations were made of the elevation of the membrane along a radius of the cylinder at various stages in the life of the membrane. Figure 3 shows the results of this test when the apparatus was tilted about 3° so that the membrane could drain. The maximum deviation from theoretical computations was found to be 2 percent with an average of less than 1 percent.

As an example of the application of the method in solving typical problems, the next problem considered was that of determining the temperature distribution through a rectangular thick-walled cylinder. A model was constructed in the same manner as the one for the circular cylinder, except that in this case the outer boundary was in the form of a 6-inch square at an elevation of 0.179 inch. The temperature here was assumed to be 358°F. The inside plate was 3 inches square at an elevation of 0.521 inch, and the temperature was assumed to be 1042°F. Thus a change of 0.001 inch elevation of the membrane represented a change in the temperature of 2°F. Figure 4 shows a contour map of the isotherms obtained from measurements taken on opposite corners.

The last experiment performed was an investigation of a problem as outlined by McAdams (8). Consider an insulator 4 inches thick, bounded on one side by an

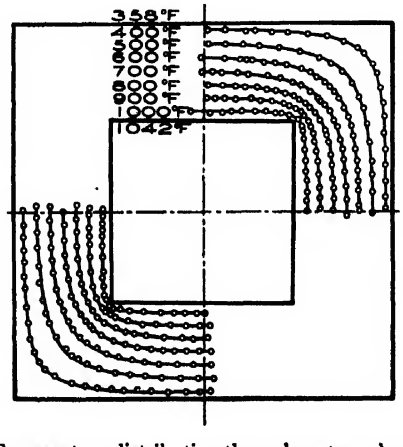


FIG. 4. Temperature distribution through rectangular thick-wall cylinder. Age of membrane 00:30 hr. to 2:00 hr. Model tilted 3° for drainage. Elevation of inside rectangle 0.521 in., outside rectangle 1.79 in.

isothermal metal plate at a temperature of 602°F, and on the lower side by another isothermal metal plate at 200°F. At intervals of 8 inches the lower part of the insulator is penetrated by metal ribs which extend 2 inches into the insulator. The ribs are also at 200°F, and there is no heat flow in the longitudinal direction of the insulator.

A model was constructed to the exact dimensions as those of the problem, except that the longitudinal dimension was made to be 4 inches since each 4-inch section would be symmetrical with the next section. On adiabatic sections of the boundary vertical plates were inserted and the membrane was allowed to seek its own level along the smooth surfaces of these plates. Figure 5 shows a plot of the isotherms obtained.

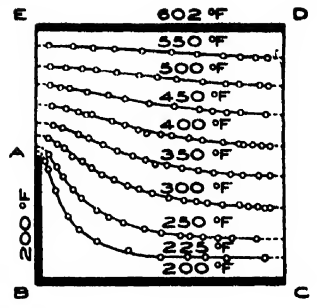


FIG. 5. Temperature distribution through an insulator. Insulator of thickness CD bounded by surface ED at a temperature of 602°F and by surface BC at a temperature of 200°F; at intervals of 2 BC the lower part of the insulator is penetrated by ribs AB also at a temperature of 200°F. Age of membrane 00:30 hr. to 2:00 hr. Model tilted 3° for drainage.

DISCUSSION OF RESULTS

From the results obtained in the investigation of the problem for which the analytical solution is known, it may be seen that under the proper conditions of test, the accuracy of the membrane analogy is well within the limits required in the solution of most heat-conduction problems. The age of the membrane produces an appreciable effect upon the results obtained, and it is recommended that the membrane be allowed to drain for at least one hour before any measurements are taken. If the measurements extend over a long period of time the membrane becomes fragile, but may be rejuvenated by placing a drop of soap solution on the probe and transferring the drop of solution to the membrane. We conclude from the results of the first tests that the optimum slope of the membrane is about 30 percent. When possible the whole apparatus should be tilted to an angle of about 3° in order that the section on which the measurements are being made may be drained of excess liquid.

In the solution of problems which require plotting of isotherms throughout the member, the greatest error to be encountered is the accuracy with which the model is constructed. Care should be taken to insure proper geometric similarity and a variation in elevation of the isothermal boundaries of less than 0.002 inch. It was found that towards the boundaries the values obtained in determining Q/k vary appreciably, but in the central region where greater accuracy is expected the values are in close agreement.

REMARKS

Although this type of solution is not suggested in competition with alternative methods, its utility lies in the solution of problems in which the geometric considerations produce tedious computations when numerical approaches are attempted. The equipment necessary for the execution of this type solution is quite simple and inexpensive as compared to the apparatus employed with electrical-geometrical analogues.

Further investigation of the membrane analogy is indicated in the realm of unsteady-state heat-conduction.

BIBLIOGRAPHY

(1) B. A. Christopherson and R. V. Southwell, "Relaxations methods applied to engineering problems 111. Problems involving two independent variables," Proc. Roy. Soc. **A168**, 31–350 (1938).

(2) H. W. Emmons, "The numerical solution of heat-conduction problems," Trans. A.S.M.E. **63**, No. 6, 607-615 (August, 1943).

(3) A. A. Griffith and G. I. Taylor, "The use of soap films in solving torsion problems," Adv. Com. Aeronautics, Tech. Report (British) No. 333, **3**, 920–937 (1917–1918).

(4) A. A. Griffith and G. I. Taylor, "The determination of the torsional stiffness and strength of cylindrical bars of any shape," Adv. Com. Aeronautics, Tech. Report (British) No. 334, **3**, 910–920 (1917–1918).

(5) A. A. Griffith and G. I. Taylor, "The application of soap films to the determination of the torsion and flexure of hollow shafts," Adv. Com. Aeronautics, Tech. Report (British) No. 392, **3**, 938–949 (1917–1918).

(6) A. A. Griffith and G. I. Taylor, "The problem of flexure and its solution by the soap-film method," Adv. Com. Aeronautics, Tech. Report (British) No. 399, **3**, 950–969 (1917–1918).

(7) C. F. Kayan, "An electrical geometrical analogue for complex heat flow," Trans. A.S.M.E. **67**, No. 8, 713–718 (November, 1945).

(8) W. H. McAdams, *Heat Transmission* (McGraw-Hill Book Company, Inc., New York, 1942), second edition, p. 1–25.

(9) A. J. Miles and E. A. Stephenson, "Pressure distribution in oil and gas reservoirs by membrane analogy," Trans. A.I.M.E. Tech. Pub., No. 919 (May, 1938).

(10) T. S. Nickerson and G. M. Dusinberre, "Heat transfer through thick insulation on cylindrical enclosures," Trans. A.S.M.E. Paper No. 47-A-63 (December, 1947).

(11) T. P. Neubauer and O. W. Boston, "Torsional stress analysis of twist-drill sections by membrane analogy," Trans. A.S.M.E. **69**, No. 8, 897–902 (November, 1947).

(12) V. Paschkis and H. D. Baker, "A method for determining unsteady-state heat transfer by means of an electrical analogy," Trans. A.S.M.E. **64**, No. 2, 105–112 (February, 1942).

(13) S. Timoshenko, *Theory of Elasticity* (McGraw-Hill Book Company, Inc., New York, 1934), pp. 239–242, 258–263.

(14) G. W. Trayer and H. W. March, "The torsion of members having sections common in aircraft construction," Nat. Ave. Comm. Aeronautics, Tech. Report No. 334, 675–705 (1930).

Work Functions of the Elements

HERBERT B. MICHAELSON
Product Development Laboratories, Sylvania Electric Products, Inc., Kew Gardens, New York
(Received September 6, 1949)

Like the chemical properties of the elements, the work function ϕ is a periodic function of atomic number, as shown clearly by a new compilation of published data for 57 elements. The progressive rise and fall of work function values throughout the table of the elements appears to be sufficiently regular to permit approximations to be made of values for the metallic and semimetallic elements on which no data has yet been published. Since the first ionization potential E_i and electrode potential E^0 exhibit a similar periodic function of atomic number, a striking resemblance exists among the respective plots of ϕ, E_i, and E^0.

I. INTRODUCTION

ALTHOUGH the published data on the work functions of metallic and semiconductor elements is by no means complete, the large number of measurements that have appeared in the literature indicate a well-defined periodicity of the values in the table of the elements. Several attempts have been made in recent years to establish relationships between work function and other periodic physical properties. Empirical equations of this type may provide a criterion for accurate values of work function, provided, of course, that they are derived from reliable data. Since, however, measurements of work function are extremely sensitive to the presence of surface impurities, particularly oxides and gases, there has been a rather large range of published values for each element. In addition, any comparison of published data is further complicated by the anisotropy[1,2] allotropy[3,4] and temperature dependence[5–8] of work function. Regardless of the difficulties in selecting a precise value of the work function for each element, the periodicity in the table of the elements can be shown clearly if *representative* values are cited for a large number of elements. Previously published lists of work functions have given data for less than half of the known elements,[9–16] while the present compilation provides values for 57 elements.

Most of the published methods of calculating work function ϕ have dealt with the relation to physical properties. The empirical formula of Rother and Bomke[17] expresses ϕ in terms of density, atomic weight, and electronic valence. According to Chittum[18] and Gombas,[19] ϕ is a function of the lattice energy in the crystal. Bartelink[20] derived an approximate relationship between work function and the reciprocal of atomic volume for the alkali metals. Another equation was developed by Wigner and Bardeen[21,22] for work functions of the univalent metals in terms of the heats of sublimation.

In addition, there have been a few attempts to correlate ϕ with other electrical measurements. Langmuir[23] observed for nine metals a general parallelism between thermionic work function and electrode potential E^0. This similarity was also noted by Scarpa[24] for 16 metals. More recently an interesting inverse relationship between hydrogen overpotential and ϕ was reported by Bockris.[25] In addition, Bedreag[26] has pointed out that for the alkali metals the first ionization potential E_i is equal to twice the work function. The purpose of this paper is to present a list of the work function data published during the past twenty-five years that have come to the writer's attention and to show graphically an apparent relationship between work function and other electrical data when plotted as a function of atomic number.

II. PERIODICITY OF WORK FUNCTION IN THE TABLE OF THE ELEMENTS

The relationship between work function and atomic number has been widely recognized in the literature.

[1] R. Smoluchowski, Phys. Rev. **60**, 661 (1941).
[2] S. T. Martin, Phys. Rev. **56**, 947 (1939).
[3] H. B. Wahlin, Phys. Rev. **61**, 509 (1942).
[4] A. Goetz, Phys. Rev. **33**, 373 (1929).
[5] S. Seely, Phys. Rev. **59**, 75 (1941).
[6] J. G. Potter, Phys. Rev. **58**, 622 (1940).
[7] A. H. Smith, Phys. Rev. **75**, 953 (1949).
[8] J. J. Markham and P. H. Miller, Jr., Phys. Rev. **75**, 959 (1949).
[9] S. Dushman, "Thermal emission of electrons," *International Critical Tables* (McGraw-Hill Book Company, Inc., New York, 1929), Vol. VI, pp. 53–4, data for 15 elements.
[10] S. Dushman, Rev. Mod. Phys. **2**, 381 (1930), data for 26 elements.
[11] A. L. Reimann, *Thermionic Emission* (John Wiley and Sons, Inc., New York, 1934), pp. 72–102, data for 32 elements.
[12] A. L. Hughes and L. A. DuBridge, *Photoelectric Phenomena* (McGraw-Hill Book Company, Inc., New York, 1932), pp. 65–6, data for 32 elements.
[13] O. Klein and E. Lange, Zeits. f. Elektrochemie **43**, 570 (1937), data for 43 elements.
[14] J. A. Becker, Rev. Mod. Phys. **7**, 123 (1935), data for 40 elements.
[15] Landolf-Boernstein, *Physikalisch-Chemische Tabellen* (1936), data for 37 elements.
[16] C. Herring and M. H. Nichols, Rev. Mod. Phys. **21**, 185 (1949), data for 23 elements (1935–1949).
[17] F. Rother and H. Bomke, Zeits. f. Physik **86**, 231 (1933).
[18] J. F. Chittum, J. Phys. Chem. **38**, 79 (1934).
[19] P. Gombas, Nature **157**, 668 (1946); Muegyetemi Koezlemenyek **1**, 25 (1947); Hung. Acta Phys. **1**, 1 (1947).
[20] E. H. B. Bartelink, Physica **3**, 193 (1936).
[21] E. Wigner and J. Bardeen, Phys. Rev. **48**, 84 (1935).
[22] J. Bardeen, Phys. Rev. **49**, 653 (1936).
[23] I. Langmuir, Trans. Amer. Electrochem. Soc. **29**, 125 (1916).
[24] O. Scarpa, Accad, Naz. d. Lincei, Rend. Cl. Sci. Fis. Nat., Atti **2**, 1062 (1941); Nuovo Cimento **17**, 54 (1940).
[25] J. O'M. Bockris, Nature **159**, 539 (1947); Trans. Faraday Soc. **43**, 417 (1947).
[26] C. G. Bedreag, Comptes Rendus **223**, 354 (1946).

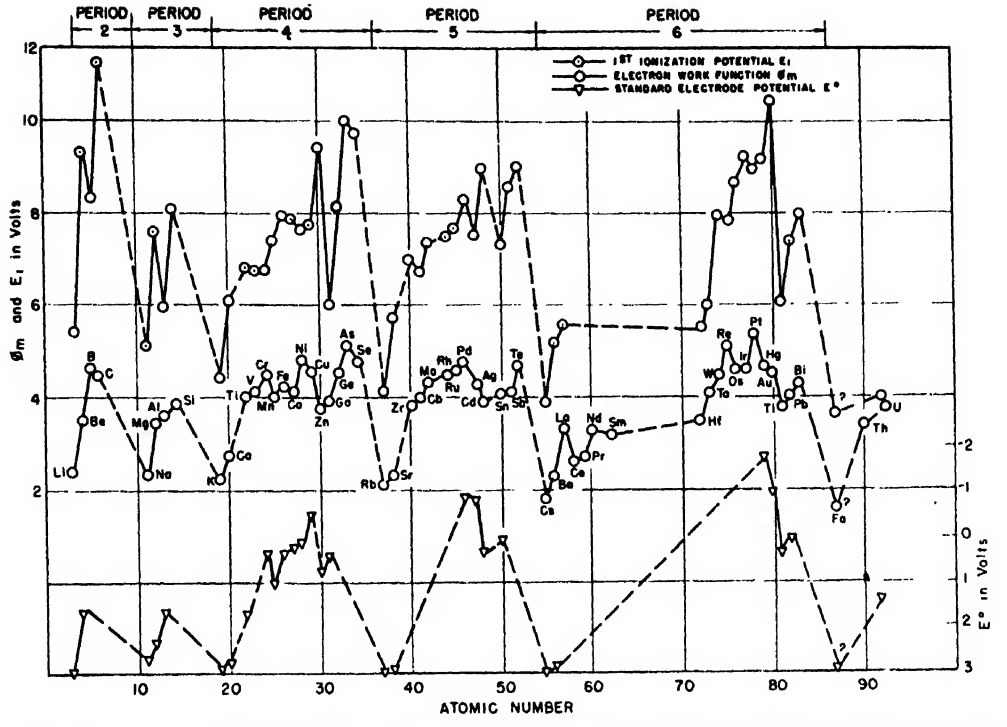


FIG. 1. Plots of work function, first ionization potential, and standard electrode potential *versus* atomic number.

Rother and Bomke,[17] Morecroft,[27] Klein and Lange,[28] and Scarpa[24] have all shown plots of ϕ as a function either of atomic number or atomic weight, the general shape of the curve being the same in either case. In Fig. 1 this previous work has been extended to include published data for a considerably larger number of elements. The data for work functions are taken from the appendix, which lists measurements published during the period 1924–1949. For most of the elements, values obtained by thermionic, photoelectric, or contact potential methods are about the same, and, to obtain *representative* values of work function, the unweighted mean ϕ_m of all data for each element was used for the plot in Fig. 1. For a few of the elements, some definite variance appears to exist among values obtained by different methods of measurement. Examples are the well-known discrepancies between photoelectric thresholds and thermionic work functions for the semiconductor elements.[29,30] These discrepancies, however, are not great enough to invalidate the representative values ϕ_m given here.

Since the temperature variation of ϕ for metals is quite small, the differences in temperature for the various measurements cited in the appendix are also of little consequence in considering the relationship between work function and atomic number. For example, the variation of ϕ with temperature for tungsten is only 3.4×10^{-5} ev/degree, according to Seely.[5]

The curve for ϕ_m in Fig. 1 exhibits a fairly regular rise and fall as it passes through each period in the table of the elements. When viewed from left to right on the periodic chart in Table I, the values in each period show a tendency to rise, except in the second-half of the periods 4, 5, and 6, where the elements in groups of *IB*, *IIB*, and *IIIB* have progressively lower work functions. The peculiar regularities in the various groups and periods are similar to the behavior of chemical and physical properties. A striking resemblance among the curves for ϕ, E_i, and E^0 is in evidence in Fig. 1, where the ionization potentials are from Hubbard and Meggers[31] and the electrode potentials* from Latimer and Hildebrand.[32]

The periodicity of work function in Fig. 1 and Table I appear to be sufficient to provide a basis for the prediction of values for elements on which no data has yet

[27] J. H. Morecroft, *Electron Tubes and Their Applications* (John Wiley and Sons, Inc., New York, 1936), p. 39.
[28] O. Klein and E. Lange, Zeits. f. Elektrochemie 44, 542 (1938).
[29] Apker, Taft, and Dickey, Phys. Rev. 74, 1462 (1948).
[30] E. U. Condon, Phys. Rev. 54, 1089 (1938).

[31] H. D. Hubbard and W. F. Meggers, *Key to Periodic Chart of the Atoms* (W. M. Welch Scientific Company, Chicago, 1947), pp. 22–5.

* The values of E^0 cited in this paper are those of the couple between the zero oxidation state and the next highest known state. For example, the couple for iron is

$$Fe \rightarrow Fe^{++} 2e^-, \quad E^0 = +0.440 \text{ volt.}$$

The scale for E^0 values is based on zero volts for the hydrogen gas—hydrogen ion couple.

[32] W. M. Latimer and J. H. Hildebrand, *Reference Book of Inorganic Chemistry* (The Macmillan Company, New York, 1947), pp. 474–6.

TABLE I. Periodic chart of work functions, mean of published data, 1924–1949.

Group	I A	I B	II A	II B	III A	III B	IV A	IV B	V A	V B	VI A	VI B	VII A	VII B	VIII		
Period 2	Li 2.39			Be 3.37		B 4.5		C 4.39									
3	Na 2.27			Mg 3.46		Al 3.74		Si 4.1		P —		S —					
4	K 2.15		Ca 2.76		Sc —		Ti 4.09		V 4.11		Cr 4.51		Mn 3.95		Fe 4.36	Co 4.18	Ni 4.84
		Cu 4.47		Zn 3.74		Ga 3.96		Ge 4.56		As 5.11		Se 4.72					
5	Rb 2.13		Sr 2.35		Y —		Zr 3.84		Cb 3.99		Mo 4.27		Tc —		Ru 4.52	Rh 4.65	Pd 4.82
		Ag 4.28		Cd 3.92		In —		Sn 4.11		Sb 4.08		Te 4.73					
6	Cs 1.89		Ba 2.29		La 3.3		Hf 3.53		Ta 4.12		W 4.50		Re 5.1		Os 4.55	Ir 4.57	Pt 5.29
		Au 4.58		Hg 4.52		Tl 3.76		Pb 4.02		Bi 4.28		Po —					
7	Fa —		Ra —		Ac —		Th 3.41		Pa —		U 3.74						
Rare earths	Ce 2.7	Pr 2.7	Nd 3.3	Sm 3.2													

appeared in the literature:

Element	Atomic number	ϕ (hypothetical)
Sc	21	3.3 ev
Y	39	3.3
Tc	43	4.4
In	49	4.0
Po	84	4.6
Fa	87	1.5
Pa	91	3.3

If these seven elements follow the periodic electrical properties exhibited by the 57 elements in Fig. 1, the predicted work functions should be correct within a few tenths of a volt.

The writer wishes to express his appreciation to Dr. Harold Jacobs,† George Hees and Henry W. Parker, of the Central Engineering Laboratories, for many enlightening discussions.

APPENDIX.‡ ELECTRON WORK FUNCTIONS OF THE ELEMENTS

(Published Data, 1924-1949)

Element	Mean value ϕ_m	Thermionic work function	Photoelectric work function	Work function by contact potential method	Reference
Ag	4.28	3.09			(1)
		3.56			(2)
		4.08			(3)
			3.67		(4)
			4.1 to 4.75		(5)
			4.56 (600°)		(6)
			4.73 (20°)		(6)
			4.75 (111) face		(7)
			4.81 (100) face		(7)
				4.33	(8)
				4.44	(9)
				4.79 (100) face	(10)
Al	3.74		2.98		(4)
			3.43		(11)
			4.08		(12)
			4.20		(13)
			4.36		(14)
				3.38	(9)
As	5.11		5.11		(15)
Au	4.58	4.0 to 4.58			(2)
		4.32			(3)
			4.73 (740°C)		(16)
			4.82 (20°C)		(16)
			4.86 to 4.92		(5)
				4.46	(9)
B	4.5		4.4 to 4.6*		(17)
Ba	2.29	2.11			(18)
				1.73	(9)
				2.39	(19)
			2.48		(20)
			2.49		(20)
			2.52		(22)
Be	3.37		3.17		(15)
			3.30		(23)
			3.92		(24)
				3.10	(9)
Bi	4.28		4.14		(11)
			4.22 to 4.25		(25)
			4.31		(26)
			4.44		(23)
			4.46		(27)
				4.17	(9)

† Now at Evans Signal Laboratory, Thermionics Branch, Belmar, New Jersey.
‡ Values indicated by an asterisk (*) were not new determinations, but were obtained as a check on experimental conditions.

Element	Mean value ϕ_m	Thermionic work function	Photoelectric work function	Work function by contact potential method	References[a]
C	4.39	4.00			(28)
		4.34			(29)
		4.39			(30)
			4.81		(31)
Ca	2.76	2.24			(32)
			2.42		(15)
			2.706		(22)
			2.76		(33)
			3.20		(34)
			3.21		(35)
Cb	3.99	3.96			(36)
		4.01			(37)
Cd	3.92		3.68		(15)
			3.73		(4)
			3.94		(11)
			4.07		(38)
			4.09[b]		(13)
				4.00	(9)
Ce*	2.7	2.6			(39)
				2.81	(34)
Co	4.18	4.40			(40)
			3.90		(33)
			4.12 to 4.25		(41)
				4.21	(9)
Cr	4.51	4.60			(42)
		4.7			(43)
			4.37		(26)
				4.38	(9)
Cs	1.89	1.81			(44)
			1.9		(45)
			1.96		(46)
Cu	4.47	3.85			(1)
		4.26			(2)
		4.38			(3)
		4.55[b]			(47)
			4.07		(4)
			4.18		(33)
			4.86 (111) face		(48)
			5.61 (100) face		(48)
				4.46	(9) (49)
Fe	4.36	4.04			(50)
		γ4.23			(40)
		β4.48			(40)
		4.77			(51)
			3.91		(33)
			3.92		(55)
			4.72		(52)
			4.77		(53)
				4.40	(9)
Ga	3.96		4.12		(15)
				3.80	(9)
Ge	4.56		1.29		(33)
			4.5*		(54)
			4.73		(15)
			4.80		(18)
				4.50	(9)
Hf	3.53	3.53			(56)
					(57)
Hg	4.52		4.50		(58)
			4.52		(59) (60)
			4.53		
Ir	4.57			4.57	(9)
K	2.15				(61)
			2.0		(23)
			2.12		(45, 46)
			2.24		(62)
			2.26		
La	3.3	3.3			(39)
					(45)
Li	2.39		2.28		(15)
			2.42		(63)
				2.49	
Mg	3.46		2.74		(15)
			<3.0		(64)
			3.59		(65)
			3.62		(66)
			3.68		(24)
			3.79		(67)
				3.58	(9)
				3.78	(68)

Element	Mean value ϕ_m	Thermionic work function	Photoelectric work function	Work function by contact potential method	References[a]
Mn	3.95		3.76		(15)
				4.14	(9)
Mo	4.27	4.15			(69)
		4.17			(70)
		4.19			(71)
		4.20			(72)
		4.32			(73)
		4.33			(74)
		4.38			(75)
		4.44			(76)
			4.15		(69)
			4.34		(26)
				4.08	(77)
				4.48	(9)
Na	2.27		2.06		(15)
			2.25		(78)
			2.28		(20)
			2.29		(24)
			2.47		(45)
				2.26	(79)
Nd	3.3	3.3			(39)
Ni	4.84	4.61			(40)
		4.63			(50)
		5.03			(80)
		5.1[b]			(47)
		5.24			(116)
			4.06		(32)
			4.87		(26)
			5.01		(53)
			5.05 (625°K)		(116)
			5.20 (1180°K)		(116)
				4.32	(9)
				4.96 (300°K)	(81)
Os	4.55			4.55	(9)
Pb	4.02		3.97		(4)
			4.14		(11)
				3.94	(9)
Pd	4.82	4.99			(82)
			4.97		(82)
				4.49	(9)
Pr	2.7	2.7			(39)
Pt	5.29	5.08			(28)
		5.29			(83)
		5.32			(84)
		6.27			(85)
			4.09		(4)
			6.35		(86)
				4.52	(9)
				5.36	(87)
Rb	2.13		2.09		(46)
			2.16		(45)
Re	5.1	5.1			(88)
			~5.0		(89)
Rh	4.65	4.58			(90)
		4.80			(91)
			4.57		(90)
Ru	4.52			4.52	(9)
Sb	4.08		4.01		(92)
				4.14	(9)
Se	4.72		4.62		(11)
			5.11		(15)
				4.42	(9)
Si	4.1	3.59			(30)
				4.2*	(93)
				4.5 (250°C)*	(54)
Sm	3.2	3.2			(39)
Sn	4.11		3.62		(4)
			3.87		(11)
			(liq.) 4.21		(94)
			γ4.38		(94)
			β4.50		(94)
				4.09	(9)
Sr	2.35		2.06		(95)
			2.24		(15)
			2.74		(68)

APPENDIX—*Continued*

Element	Mean value ϕ_m	Thermionic work function	Photoelectric work function	Work function by contact potential method	Reference[a]
Ta	4.12	4.07			(76)
		4.10			(96)
		4.19			(97)
			4.12		(34)
			4.16		(98)
				3.96	(9)
				4.25	(99)
Te	4.73		4.76*		(17)
				4.70	(9)
Th	3.41	3.35			(75)
			3.38		(34)
			3.47		(100)
			3.57		(11)
				3.46	(9)
Tl	4.09		3.95		(15)
			4.17		(100)
				4.14	(9)
Tl	3.76			3.84	(9)
		3.68			(101)
U	3.74	3.27			(102)
			3.63		(34)
				4.32	(9)
V	4.11		3.77		(15)
				4.44	(9)
W	4.50	4.25			(103)
		4.39 (111) face			(115)
		4.39 (116) face			(115)
		4.45			(117)
		4.46			(104)
		4.52			(28, 105)
		4.53			(106)
		4.56 (001) face			(115)
		4.58			(73, 75)
		4.68 (110) face			(115)
		4.69 (112) face			(115)
			4.35 (310) face		(107)
			4.50 (211) face		(107)
			4.54		(108)
			4.60		(34)
				4.38	(9)
Zn	3.74		3.08		(4)
			3.28		(109)
			3.32 and 3.57		(110)
			3.60		(11)
			3.89		(33)[b]
			4.24		(111)
			4.26 (0001) face		(112)
			4.307		(13)[b]
				3.40	(113)
				3.66	(9)
				4.28	(114)
Zr	3.84	4.12			(56)
			3.73		(34)
			3.91		(102)
				3.60	(9)

[a] Reference numbers:

(1) A. Wehnelt and S. Seliger, Zeits. f. Physik **38**, 443 (1926).
(2) I. Ameiser, Zeits. f. Physik **69**, 111 (1931).
(3) A. Goetz, Zeits. f. Physik **43**, 531 (1927).
(4) P. Lukirsky and S. Prilesaev, Zeits. f. Physik **49**, 236 (1928).
(5) R. H. Fowler, Phys. Rev. **38**, 45 (1931).
(6) R. P. Winch, Phys. Rev. **37**, 1269 (1931).
(7) H. E. Farnsworth and R. P. Winch, Phys. Rev. **58**, 812 (1940).
(8) P. A. Anderson, Phys. Rev. **49**, 320 (1936).
(9) O. Klein and E. Lange, Zeits. f. Elektrochemie **44**, 542 (1938).
(10) P. A. Anderson, Phys. Rev. **59**, 1034 (1941).
(11) R. Hamer, Jour. Opt. Soc. Amer. **9**, 251 (1924).
(12) J. Brady and P. Jakobsmeyer, Phys. Rev. **49**, 670 (1936).
(13) R. Suhrmann and J. Pietrzyk, Zeits. f. Physik **122**, 600 (1944).
(14) E. Gaviola and J. Strong, Phys. Rev. **49**, 441 (1936).
(15) R. Schulze, Zeits. f. Physik **92**, 212 (1934).
(16) L. W. Morris, Phys. Rev. **37**, 1263 (1931).
(17) L. Apker, E. Taft, and J. Dickey, Phys. Rev. **74**, 1462 (1948).
(18) A. L. Reimann, *Thermionic Emission* (Chapman and Hall, London, 1934), p. 37.
(19) P. A. Anderson, Phys. Rev. **47**, 958 (1935).
(20) R. J. Maurer, Phys. Rev. **57**, 653 (1940).
(21) R. J. Cashman and E. Bassoe, Phys. Rev. **55**, 63 (1939).
(22) N. C. Jamison and R. J. Cashman, Phys. Rev. **50**, 624 (1936).
(23) R. Suhrmann and A. Schallamach, Zeits. f. Physik **91**, 775 (1934).
(24) M. M. Mann, Jr. and L. A. DuBridge, Phys. Rev. **51**, 120 (1937).
(25) H. Jupnik, Phys. Rev. **60**, 884 (1941).
(26) H. C. Rentschler and D. E. Henry, J. Opt. Soc. Am. **26**, 30 (1936).
(27) A. H. Weber and C. J. Eisele, Phys. Rev. **59**, 473(A) (1941).
(28) S. Dushman, "Thermal emission of electrons," *International Critical Tables* (McGraw-Hill Book Company, Inc., New York, 1929), Vol. VI, pp. 53–4.
(29) A. L. Reimann, Proc. Phys. Soc. **50**, 496 (1938).
(30) A. Braun and G. Busch, Helv. Phys. Acta **20**, No. 1, 33 (1947).
(31) S. C. Roy, Proc. Roy. Soc. **A112**, 599 (1926).
(32) S. Dushman, Phys. Rev. **21**, 623 (1923).
(33) G. B. Welch, Phys. Rev. **32**, 657 (1928).
(34) Rentschler, Henry, and Smith, Rev. Sci. Inst. **3**, 794 (1932).
(35) I. Liben, Phys. Rev. **51**, 642 (1937).
(36) H. B. Wahlin and L. I. Sordahl, Phys. Rev. **45**, 886 (1934).
(37) A. L. Reimann and C. Kerr Grant, Phil. Mag. **22**, 34 (1936).
(38) H. Bomke, Ann. d. Physik **10**, 579 (1931).
(39) E. E. Schumaker and J. E. Harris, J. Am. Chem. Soc. **48**, 3108 (1926).
(40) H. B. Wahlin, Phys. Rev. **61**, 509 (1942).
(41) A. B. Cardwell, Phys. Rev. **38**, 2033 (1931).
(42) H. B. Wahlin, Phys. Rev. **73**, 1458 (1948).
(43) H. Koesters, Zeits. f. Physik **66**, 807 (1930).
(44) H. H. Kingdon, Phys. Rev. **25**, 892 (1925).
(45) A. R. Olpin, quoted by Hughes and DuBridge, *Photoelectric Phenomena* (1932).
(46) J. J. Brady, Phys. Rev. **41**, 613 (1932).
(47) W. P. Dyke, Thesis, University of Washington (1946).
(48) N. Underwood, Phys. Rev. **47**, 502 (1935).
(49) P. A. Anderson, Phys. Rev. **76**, 388 (1949).
(50) W. Distler and G. Monch, Zeits. f. Physik **84**, 271 (1933).
(51) G. Siljeholm, Ann. d. Physik **10**, 178 (1931).
(52) A. B. Cardwell, Proc. Nat. Acad. Sci. **14**, 439 (1928).
(53) G. N. Glasoe, Phys. Rev. **38**, 1490 (1931).
(54) A. H. Smith, Phys. Rev. **75**, 953 (1949).
(55) G. B. Welch, Phys. Rev. **31**, 709(A) (1928).
(56) C. Zwikker, Physik. Zeits. **30**, 578 (1929).
(57) H. Cassel and A. Schneider, Naturwiss **22**, 464 (1934).
(58) Roller, Jordan, and Woodward, Phys. Rev. **38**, 396 (1931).
(59) C. B. Kazda, Phys. Rev. **26**, 643 (1925).
(60) W. B. Hales, Phys. Rev. **32**, 950 (1928).
(61) H. E. Ives, J. Opt. Soc. Am. **8**, 551 (1924).
(62) H. Mayer, Ann. d. Physik **29**, 129 (1937).
(63) P. A. Anderson, Phys. Rev. **75**, 1205 (1949).
(64) C. Kenty, Phys. Rev. **43**, 776(A) (1933).
(65) J. Cashman and S. Huxford, Phys. Rev. **48**, 734 (1935).
(66) R. J. Cashman and E. Bassoe, Phys. Rev. **53**, 919(A) (1938).
(67) R. J. Cashman, Phys. Rev. **54**, 971 (1938).
(68) P. A. Anderson, Phys. Rev. **54**, 753 (1938).
(69) L. A. DuBridge and W. W. Roehr, Phys. Rev. **42**, 52 (1932).
(70) H. B. Wahlin and J. A. Reynolds, Phys. Rev. **48**, 751 (1935).
(71) H. Grover, Phys. Rev. **52**, 982 (1937).
(72) R. W. Wright, Phys. Rev. **60**, 465 (1941).
(73) A. J. Ahearn, Phys. Rev. **44**, 277 (1933).
(74) E. Freitag and F. Kruger, Ann. d. Physik **21**, 697 (1934).
(75) C. Zwikker, Proc. Roy. Acad. Amsterdam **29**, 792 (1926).
(76) Dushman, Rowe, Ewald, and Kidner, Phys. Rev. **25**, 338 (1925).
(77) C. W. Oatley, Proc. Roy. Soc. **A155**, 218 (1936).
(78) Z. Berkes, Math. Phys. Lapok **41**, 131 (1934).
(79) E. Patai, Zeits. f. Physik **59**, 697 (1930).
(80) G. W. Fox and R. M. Bowie, Phys. Rev. **44**, 345 (1933).
(81) R. C. L. Bosworth, Trans. Faraday Soc. **35**, 397 (1939).
(82) L. A. DuBridge and W. W. Roehr, Phys. Rev. **39**, 99 (1932).
(83) H. L. Van Velzer, Phys. Rev. **44**, 831 (1933).
(84) L. V. Whitney, Phys. Rev. **50**, 1154 (1936).
(85) L. A. DuBridge, Phys. Rev. **32**, 961 (1928).
(86) L. A. DuBridge, Phys. Rev. **31**, 236 (1928).
(87) C. W. Oatley, Proc. Phys. Soc. **51**, 318 (1939).
(88) Agte, Alterthum, Becker, Heyne, and Moers, Naturwiss. **19**, 108(1931).
(89) A. Eagelmann, Ann. d. Physik **17**, 185 (1933).
(90) E. H. Dixon, Phys. Rev. **37**, 60 (1931).
(91) H. B. Wahlin and L. V. Whitney, J. Chem. Phys. **6**, 594 (1938).
(92) V. Middel, Zeits. f. Physik **105**, 358 (1937).
(93) W. B. Meyerhof, Phys. Rev. **71**, 727 (1947).
(94) A. Goetz, Phys. Rev. **33**, 373 (1929).
(95) R. Doepel, Zeits. f. Physik **33**, 237 (1925).
(96) A. B. Cardwell, Phys. Rev. **47**, 628 (1935).
(97) M. D. Fiske, Phys. Rev. **61**, 513 (1942).
(98) A. B. Cardwell, Phys. Rev. **38**, 2041 (1931).
(99) W. Heinze, Zeits. f. Physik **109**, 459 (1938).
(100) H. C. Rentschler and D. E. Henry, Trans. Electrochem. Soc. **87**, 289 (1945–6).
(101) R. Suhrmann and H. Csesch, Zeits. Chem. **B28**, 215 (1935).
(102) W. L. Hole and R. W. Wright, Phys. Rev. **56**, 785 (1939).
(103) A. H. Warner, Proc. Nat. Acad. Sci. **13**, 56 (1927).
(104) G. M. Fleming and J. E. Henderson, Phys. Rev. **58**, 887 (1940).
(105) W. B. Nottingham, Phys. Rev. **47**, 806(A) (1935).
(106) H. Freitag and F. Krueger, Ann. d. Physik **21**, 697 (1934).
(107) C. E. Mendenhall and C. F. DeVoe, Phys. Rev. **51**, 346 (1937).
(108) A. H. Warner, Phys. Rev. **38**, 1871 (1931).
(109) A. Nitsche, Ann. d. Physik **14**, 463 (1932).
(110) J. H. Dillon, Phys. Rev. **38**, 408 (1931).
(111) C. F. DeVoe, Phys. Rev. **50**, 481 (1936).
(112) W. Klug and H. Steyskal, Zeits. f. Physik **116**, 415 (1940).
(113) C. W. Oatley, Proc. Roy. Soc. **A155**, 218 (1936).
(114) P. A. Anderson, Phys. Rev. **57**, 122 (1940).
(115) C. Herring and M. H. Nichols, Rev. Mod. Phys. **21**, 185 (1949).
(116) A. B. Cardwell, Phys. Rev. **76**, 125 (1949).
(117) M. H. Nichols, Phys. Rev. **59**, 944(A) (1941).

[b] By field current method.

Compilation of Body Wave Velocity Data for Cubic and Hexagonal Metals

LOUIS GOLD
Brown University, Providence, Rhode Island
(Received October 14, 1949)

The directional variation of body wave velocities has been computed for cubic and hexagonal metals where stiffness coefficients are known; the essential data is presented in the form of tables and graphs. The results for aluminum, copper, gold, iron(α), lead and silver are tabulated in terms of the directional indices, and a stereographic coordinate system has been devised for presenting these data in the form of iso-velocity contours for the three wave vectors. The compressional and shear velocities for cadmium, magnesium, and zinc are plotted as a function of the angle formed by the direction of wave propagation and the hexagonal (c) axis from which have been derived the theoretical average velocities which may be compared with measured values for polycrystalline specimens.

1. INTRODUCTION

THE present report makes available for the first time a systematic summary of the velocity of body waves in a number of cubic and hexagonal metals for which the stiffness coefficients have been determined. Recent interest in ultrasonic wave propagation in solids,[1] particularly the question of the scattering of sound waves in polycrystalline metals,[2] has served to stimulate an orderly evaluation and presentation of such velocity data as are offered here. It is not unlikely that the tables and graphs will find application in delay-line design work, quantitative treatment of sound attenuation in polycrystalline aggregates, etc.

Two distinct graphical schemes were devised: a stereographic coordinate system is employed for the cubic metals, whereas a simple plot of velocity *versus* the angle defined by the hexagonal axis and the direction of wave propagation suffices for the hexagonal metals. The crystal symmetry of the hexagonal-close packed metals is such as to effect considerable reduction in the involved relation so that calculation of an average velocity is feasible; this is not true for the cubic metals.

Only those aspects of crystal elasticity theory directly pertinent to the details of the calculations are given. Adequate accounts are available for those who desire further elaboration.[3]

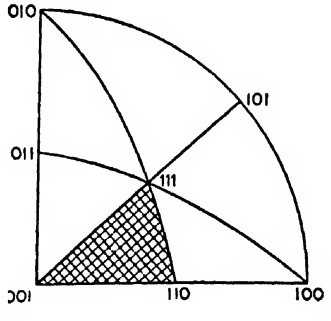


FIG. 1. Quadrant of standard projection for cubic crystal showing the six equivalent triangles of the type indicated by shaded area.

2. TREATMENT OF CUBIC METALS

The three roots of the equation

$$C_n^3 - I_1 C_n^2 + I_2 C_n - I_3 = 0 \qquad (1)$$

are related to the wave velocities by the familiar expression $v=(C_n/\rho)^{\frac{1}{2}}$. The coefficients I_1, I_2, and I_3 for crystals of any type of symmetry are defined below

$$\begin{aligned} I_1 &= k_{11}+k_{22}+k_{33} \\ I_2 &= k_{11}k_{22}+k_{22}k_{33}+k_{11}k_{33}-(k_{12}^2+k_{23}^2+k_{13}^2) \\ I_3 &= k_{11}k_{22}k_{33}+2k_{12}k_{13}k_{23}-(k_{12}^2k_{33}+k_{13}^2k_{22}+k_{23}^2k_{11}). \end{aligned} \qquad (2)$$

For crystals of cubic symmetry the k's are of the following nature . . .

$$\begin{aligned} k_{11} &= c_{11}\alpha^2+c_{44}(\beta^2+\gamma^2) \\ k_{22} &= c_{11}\beta^2+c_{44}(\alpha^2+\gamma^2) \\ k_{33} &= c_{11}\gamma^2+c_{44}(\alpha^2+\beta^2) \\ k_{12} &= (c_{44}+c_{12})\alpha\beta \\ k_{13} &= (c_{44}+c_{12})\alpha\gamma \\ k_{23} &= (c_{44}+c_{12})\beta\gamma, \end{aligned} \qquad (3)$$

where α, β, and γ are direction cosines.

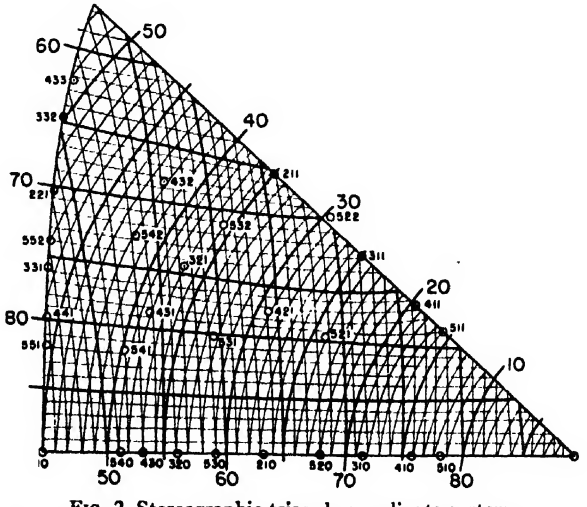


FIG. 2. Stereographic triangle coordinate system.

[1] S. Sokolov, "Absorption of Ultrasonics by Solid Bodies," Akad. Nauk. SSSR Doklady N.S. 59, 883 (1948); D. Arenberg, "Supersonic Solid Delay Lines," M.I.T. Rad. Lab. Report 932, April 1946.

[2] W. P. Mason and H. J. McSkimin, J. Acous. Soc. Am. 19, 464 (1947). By the same authors, J. App. Phys. 19, 940 (1948). W. Roth, J. App. Phys. 19, 901 (1948).

[3] I. Koga, "Thickness vibrations of piezoelectric oscillating crystals," Physics 3, 70 (1932).

TABLE I. Cubic metals.*

(Goens) $C_{11}=10.81$ $C_{12}=6.22$ $C_{44}=2.84$

Aluminum (a)

×10¹¹ dynes/cm²						
hkl	C_l ×10¹¹	C_{t_1} ×10¹¹	C_{t_2} ×10¹¹	V_l ×10⁵	V_{t_1} ×10⁵	V_{t_2} ×10⁵
100	10.82	2.84	2.84	6.33	3.24	3.24
110	11.36	2.30	2.30	6.49	2.92	2.92
111	11.54	2.48	2.48	6.54	3.03	3.03
210	11.17	2.49	2.49	6.43	3.03	3.03
211	11.37	2.66	2.47	6.49	3.14	3.03
221	11.46	2.68	2.36	6.52	3.15	2.96
310	11.02	2.64	2.64	6.39	3.13	3.13
311	11.17	2.74	2.59	6.43	3.19	3.10
320	11.28	2.38	2.38	6.46	2.97	2.97
321	11.37	2.73	2.40	6.49	3.18	2.98
331	11.41	2.76	2.33	6.50	3.20	2.94
332	11.50	2.60	2.40	6.53	3.10	2.98
410	10.95	2.71	2.71	6.37	3.17	3.17
411	11.05	2.78	2.67	6.40	3.21	3.14
421	11.24	2.77	2.49	6.45	3.20	3.04
430	11.32	2.34	2.34	6.48	2.94	2.94
431	11.36	2.78	2.36	6.49	3.21	2.95
510	10.90	2.76	2.76	6.36	3.19	3.19
511	10.98	2.80	2.72	6.38	3.22	3.17
520	11.09	2.57	2.57	6.41	3.09	3.09
521	11.14	2.79	2.57	6.42	3.21	3.09
530	11.25	2.41	2.41	6.45	2.99	2.99
540	11.34	2.32	2.32	6.48	2.93	2.93

$C_{11}=17.0$ $C_{12}=12.3$ $C_{44}=7.52$

Copper (b)

×10¹¹ dynes/cm²						
100	17.00	7.52	7.52	4.36	2.90	2.90
110	22.17	2.35	2.35	4.98	1.62	1.62
111	23.89	4.07	4.07	5.17	2.13	2.13
210	20.68	3.84	3.84	4.81	2.07	2.07
211	22.49	5.80	3.75	5.02	2.55	2.05
221	23.20	5.92	2.92	5.09	2.57	1.81
310	19.31	5.21	5.21	4.65	2.41	2.41
311	20.79	6.58	4.67	4.82	2.71	2.29
320	21.59	2.93	2.93	4.91	1.81	1.81
321	22.40	6.45	3.19	5.01	2.69	1.89
331	22.73	6.69	2.62	5.04	2.74	1.71
332	23.58	5.17	3.29	5.14	2.40	1.92
410	18.52	6.00	6.00	4.55	2.59	2.59
411	19.64	6.95	5.45	4.69	2.79	2.47
421	21.36	6.80	3.88	4.89	2.76	2.08
430	21.87	2.65	2.65	4.95	1.72	1.72
431	22.32	6.90	2.82	5.00	2.78	1.78
510	18.06	6.46	6.46	4.49	2.69	2.69
511	18.91	7.14	5.99	4.60	2.83	2.59
520	19.91	4.61	4.61	4.72	2.27	2.27
521	20.47	7.01	4.56	4.79	2.80	2.26
530	21.28	3.24	3.24	4.88	1.90	1.90
540	21.98	2.54	2.54	4.96	1.68	1.68

$C_{11}=19.4$ $C_{12}=16.6$ $C_{44}=4.00$

Gold (c)

×10¹¹ dynes/cm²						
100	19.40	4.00	4.00	3.17	1.44	1.44
110	22.00	1.40	1.40	3.38	0.85	0.85
111	22.87	2.27	2.27	3.44	1.08	1.08
210	21.15	2.25	2.25	3.31	1.08	1.08
211	22.08	3.13	2.19	3.38	1.27	1.07
221	22.50	3.21	1.69	3.41	1.29	0.94
310	20.43	2.97	2.97	3.25	1.24	1.24
311	21.15	3.53	2.72	3.31	1.35	1.19
320	21.66	1.74	1.74	3.35	0.95	0.95
321	22.05	3.48	1.87	3.38	1.34	0.98
331	22.27	3.60	1.54	3.40	1.37	0.89
332	22.70	2.83	1.87	3.43	1.21	0.99
410	20.05	3.35	3.35	3.22	1.32	1.32

TABLE I (*continued*).

(Goens) $C_{11}=19.4$ $C_{12}=16.6$ $C_{44}=4.00$

Gold (c)

×10¹¹ dynes/cm²						
hkl	C_l ×10¹¹	C_{t_1} ×10¹¹	C_{t_2} ×10¹¹	V_l ×10⁵	V_{t_1} ×10⁵	V_{t_2} ×10⁵
411	20.57	3.71	3.12	3.26	1.39	1.27
421	21.48	3.65	2.28	3.34	1.38	1.09
430	21.82	1.58	1.58	3.36	0.90	0.90
431	22.03	3.70	1.66	3.38	1.39	0.93
510	19.84	3.56	3.56	3.21	1.36	1.36
511	20.21	3.81	3.38	3.24	1.40	1.32
520	20.73	2.67	2.67	3.28	1.18	1.18
521	21.00	3.75	2.65	3.30	1.39	1.17
530	21.48	1.92	1.92	3.34	1.00	1.00
540	21.89	1.51	1.51	3.37	0.88	0.88

(Goens+Schmid) $C_{11}=23.7$ $C_{12}=14.1$ $C_{44}=11.6$

Iron(α) (d)

×10¹¹ dynes/cm²						
100	23.70	11.60	11.60	5.49	3.84	3.84
110	30.50	4.80	4.80	6.23	2.47	2.47
111	32.77	7.07	7.07	6.45	3.00	3.00
210	28.55	6.75	6.75	6.02	2.93	2.93
211	30.93	9.33	6.64	6.27	3.44	2.90
221	31.85	9.49	5.56	6.36	3.47	2.66
310	26.75	8.55	8.55	5.83	3.30	3.30
311	28.69	10.36	7.84	6.04	3.63	3.16
320	29.74	5.56	5.56	6.15	2.66	2.66
321	30.81	10.19	5.90	6.26	3.60	2.74
331	31.24	10.50	5.16	6.30	3.65	2.56
332	32.36	8.51	6.04	6.41	3.29	2.77
410	25.72	9.58	9.58	5.72	3.49	3.49
411	27.19	10.84	8.86	5.88	3.71	3.36
421	29.44	10.65	6.81	6.12	3.68	2.94
430	30.10	5.20	5.20	6.18	2.57	2.57
431	30.69	10.78	5.42	6.25	3.70	2.62
510	25.11	10.19	10.19	5.65	3.60	3.60
511	26.23	11.09	9.58	5.77	3.75	3.49
520	27.54	7.76	7.76	5.92	3.14	3.14
521	28.28	10.93	7.69	5.99	3.73	3.13
530	29.34	5.96	5.96	6.11	2.75	2.75
540	30.26	5.04	5.04	6.20	2.53	2.53

(Goens+Weerts) $C_{11}=4.77$ $C_{12}=4.03$ $C_{44}=1.44$

Lead (e)

×10¹¹ dynes/cm²						
100	4.77	1.44	1.44	2.05	1.13	1.13
110	5.84	0.37	0.37	2.27	0.57	0.57
111	6.20	0.73	0.73	2.34	0.80	0.80
210	5.51	0.70	0.70	2.20	0.79	0.79
211	5.89	1.08	0.68	2.28	0.98	0.77
221	6.05	1.11	0.49	2.31	0.99	0.66
310	5.22	0.99	0.99	2.14	0.93	0.93
311	5.52	1.25	0.88	2.21	1.05	0.88
320	5.71	0.50	0.50	2.24	0.66	0.66
321	5.88	1.22	0.55	2.28	1.04	0.70
331	5.95	1.27	0.43	2.29	1.06	0.61
332	6.13	0.96	0.56	2.32	0.92	0.71
410	5.06	1.15	1.15	2.11	1.01	1.01
411	5.28	1.32	1.05	2.16	1.08	0.96
421	5.65	1.29	0.71	2.23	1.07	0.79
430	5.77	0.44	0.44	2.26	0.62	0.62
431	5.86	1.31	0.47	2.27	1.08	0.65
510	4.97	1.24	1.24	2.09	1.05	1.05
511	5.13	1.36	1.16	2.13	1.10	1.01
520	5.34	0.87	0.87	2.17	0.87	0.87
521	5.46	1.34	0.86	2.19	1.09	0.87
530	5.64	0.57	0.57	2.23	0.71	0.71
540	5.80	0.41	0.41	2.26	0.60	0.60

TABLE I (*continued*).

×10¹¹ dynes/cm²		(Röhl) $C_{11}=12.0$	$C_{12}=8.97$	$C_{44}=4.36$		
		Silver				(1)
hkl	C_l ×10¹¹	C_{t_1} ×10¹¹	C_{t_2} ×10¹¹	V_l ×10⁵	V_{t_1} ×10⁵	V_{t_2} ×10⁵
100	12.00	4.36	4.36	3.38	2.04	2.04
110	14.84	1.52	1.52	3.76	1.20	1.20
111	15.79	2.46	2.46	3.88	1.53	1.53
210	13.98	2.38	2.38	3.65	1.50	1.50
211	14.99	3.41	2.32	3.78	1.80	1.49
221	15.40	3.49	1.83	3.83	1.82	1.32
310	13.21	3.15	3.15	3.55	1.73	1.73
311	14.02	3.84	2.86	3.65	1.91	1.65
320	14.51	1.85	1.85	3.72	1.33	1.33
321	14.95	3.78	1.99	3.77	1.90	1.38
331	15.15	3.91	1.66	3.80	1 93	1.26
332	15.62	3.07	2.03	3.86	1.71	1.39
410	12.78	3.58	3.58	3.49	1.85	1.85
411	13.38	4.04	3.29	3.57	1.96	1.77
421	14.35	3.96	2.41	3.70	1.94	1.51
430	14.67	1.69	1.69	3.74	1.27	1.27
431	14.91	4.03	1.79	3.77	1.96	1.30
510	12.54	3.82	3 82	3.46	1.91	1.91
511	12.99	4.15	3.58	3.52	1.99	1.85
520	13.54	2.82	2.82	3.59	1.64	1.64
521	13.85	4.08	2.79	3.63	1.97	1.63
530	14.33	2.03	2.03	3.69	1.39	1.39
540	14.74	1.62	1.62	3.75	1.24	1.24

* Velocities in 10^5 cm/sec.

One can readily appreciate the amount of computational effort required for the evaluation of the data in Table I. By proceeding step-wise with Eqs. (3) and (2), the undertaking reduces primarily to one of routine calculation of the coefficients of Eq. (1). The algebraic solution of the latter equation constitutes a distinct segment of the computing job and the final velocity results follow directly. The three parameters shown in Table I are associated with the longitudinal and two transverse waves, although it is understood that except

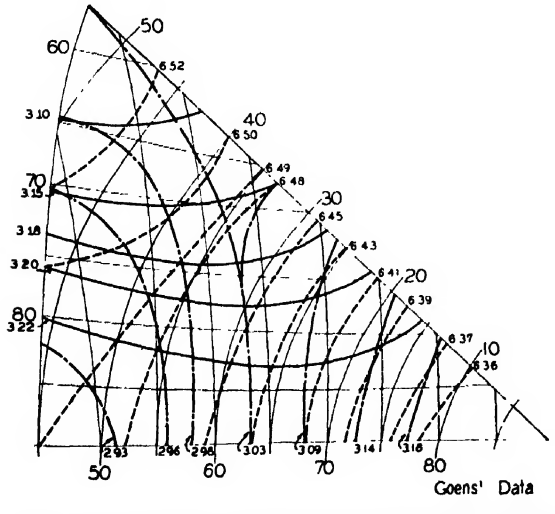


FIG. 3a. Stereographic plot of body wave velocities for aluminum showing iso-velocity contours. Velocities in 10^5 cm/sec. – – – – – – – – V_l; ———— V_{t_1}; ———— V_{t_2}.

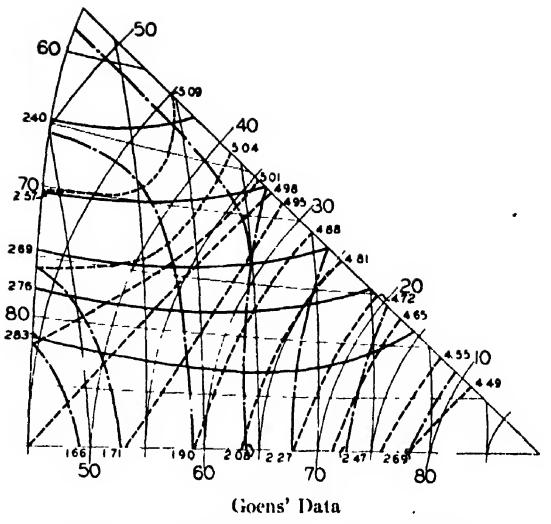


FIG. 3b. Stereographic plot of body wave velocities for copper showing iso-velocity contours. Velocities in 10^5 cm/sec. – – – – – – – – V_l; ———— V_{t_1}; ———— V_{t_2}.

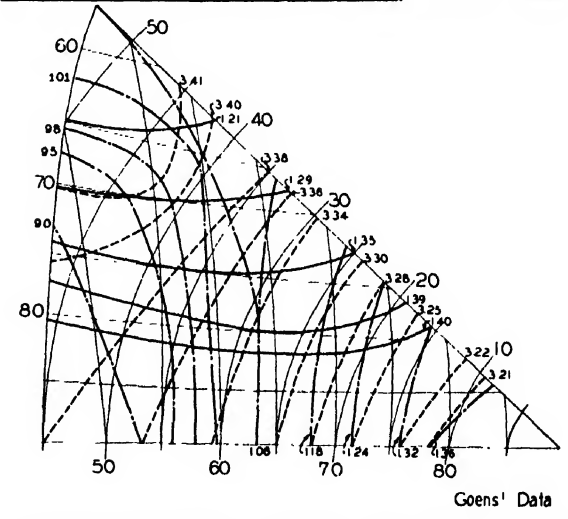
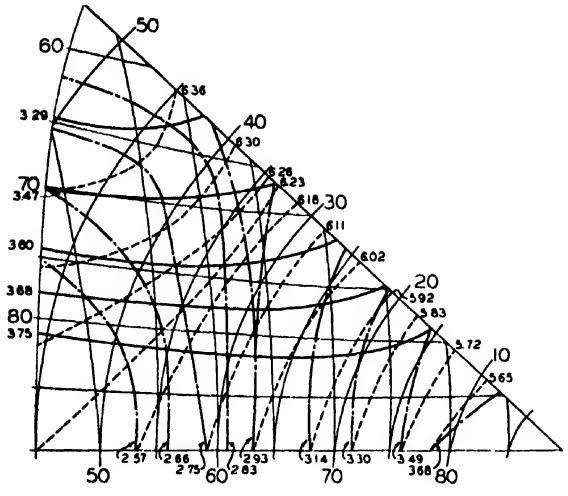


FIG. 3c. Stereographic plot of body wave velocities for gold showing iso-velocity contours. Velocities in 10^5 cm/sec. – – – – – – – – V_l; ———— V_{t_1}; ———— V_{t_2}.

for the directions 100, 110, and 111, the strict meaning of these terms does not apply.

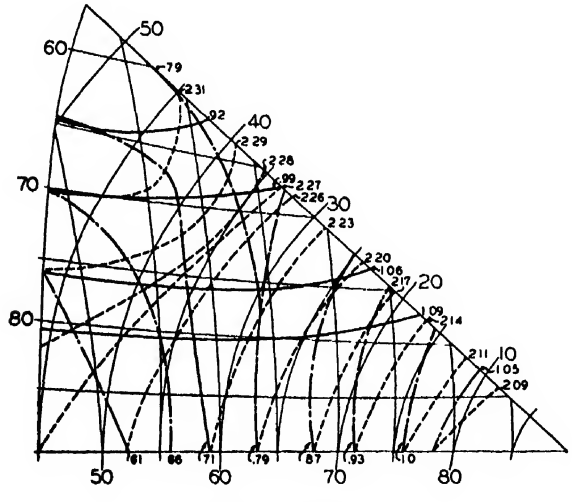
It is difficult to visualize the directional character of the body wave velocities directly from the tables given. The task of representing the data by some graphical scheme was considered. The use of a three-dimensional solid model, so popular for summarizing other directional crystalline properties, could have been invoked; but some thought was given to a two dimensional scheme, with the result that a stereographic coordinate triangle was devised for this purpose.

The underlying basis for this coordinate system may be understood in terms of the following discussion. If one considers a quadrant of a cubic crystal defined by



Goens and Schmid's Data

FIG. 3d. Stereographic plot of body wave velocities for iron showing iso-velocity contours. Velocities in 10^5 cm/sec. --------- Vt_1; —— —— Vt_1; – – – – Vt_2.

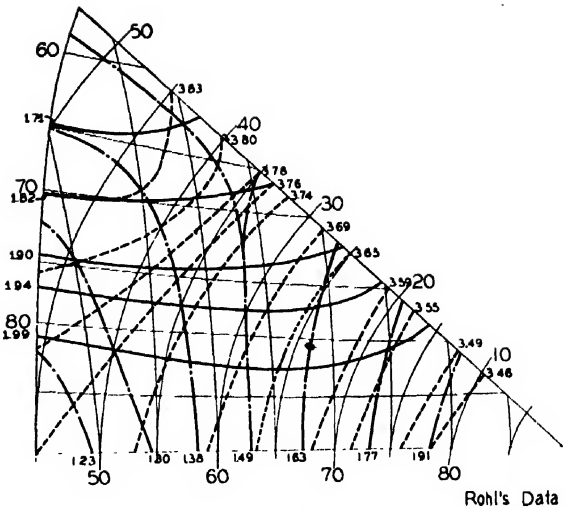


Goens and Weerts' Data

FIG. 3e. Stereographic plot of body wave velocities for lead showing iso-velocity contours. Velocities in 10^5 cm/sec. --------- Vt_1; —— Vt_1; —— —— Vt_2.

the appropriate [001] directions, then the vector along the 111 direction divides the quadrant into six equivalent regions, i.e., the directional properties are sixfold degenerate. In terms of the standard projection scheme, there are six equivalent stereographic triangles whose termini are the poles 001, 110, and 111 as shown in Fig. 1. Directional parameters such as the body wave velocities may all be plotted in one of the six triangles. To facilitate the location of each of the points within the triangle it has been found convenient to employ a coordinate system directly related to the Wulff net.

Figure 2 depicts the character of the coordinates in the stereographic triangle. The angles indicated are all



Rohl's Data

FIG. 3f. Stereographic plot of body wave velocities for silver showing iso-velocity contours. Velocities in 10^5 cm/sec. --------- Vt_1; —— —— Vt_1; —— —— —— Vt_2.

referred to the three [001] directions; the coordinate markings for two of the directions are obtained directly from the Wulff net, these being the rulings running essentially parallel to the edges defined by the 111–110 and 110–001 points. The third set of rulings is obtained by rotating the triangle through 90° and again making use of the latitude markings on the Wulff net; these lines also run along the edge 110–111, but have an opposite curvature. It may be readily verified that any point in the triangle fulfills the condition $\cos^2\alpha+\cos^2\beta$

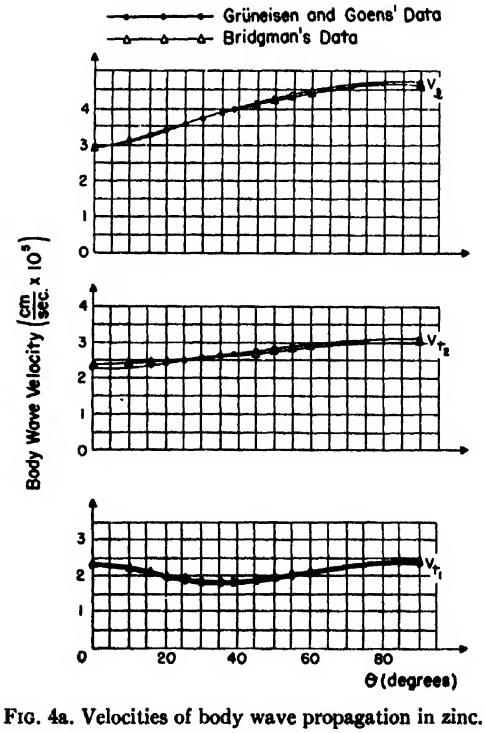


FIG. 4a. Velocities of body wave propagation in zinc.

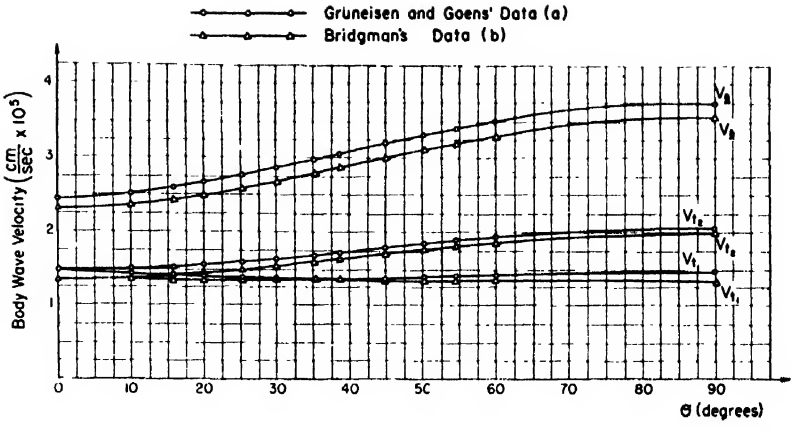


FIG. 4b. Velocities of body wave propagation in cadmium.

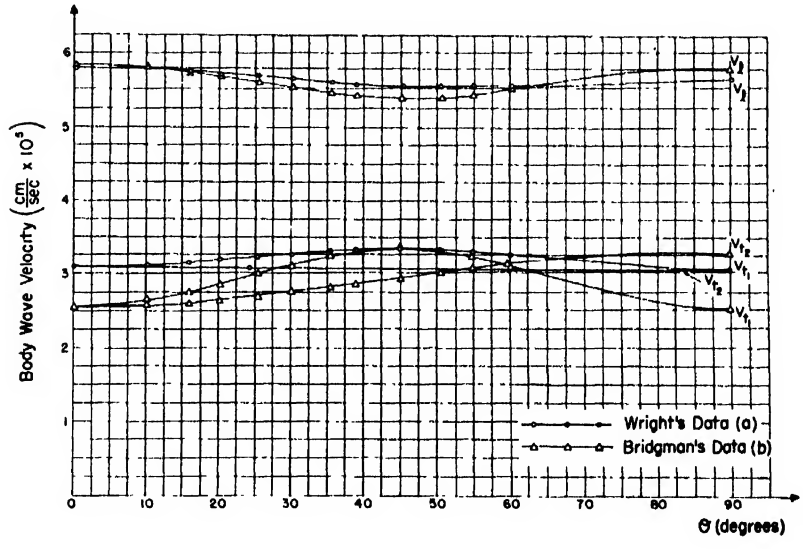


FIG. 4c. Velocities of body wave propagation in magnesium.

$+\cos^2\gamma = 1$, so that the triangle is a nomographic chart for this relation.

When the data of Table I are plotted in the stereographic triangle, it is found that lines of constant velocity can be drawn through the various points. Figures 3(a)–(f) show the common features of these iso-velocity contours for the series of metals, aluminum, copper, gold, iron, lead, and silver. One immediate application of these charts is the determination of mode separation along any given direction of propagation. Thus if one is interested in a particular compressional mode, the optimum orientation for minimizing interference from the shear modes may be inferred.

3. TREATMENT OF HEXAGONAL METALS

Hexagonal close-packed metals such as magnesium, zinc, and cadmium have symmetry which restricts the number of independent stiffness coefficients to five. It can be shown that for crystals of this kind the body wave velocities are only dependent upon the angle between the hexagonal axis and the wave propagation direction. By transforming from direction cosines to Eulerian angles the three roots of Eq. (1) become

$$\begin{aligned} C_{n1} &= \tfrac{1}{2}(c_{11}-c_{12})\sin^2\theta + c_{44}\cos^2\theta \\ C_{n2,n3} &= \tfrac{1}{2}\{(c_{11}+c_{44})\sin^2\theta + (c_{33}+c_{44})\cos^2\theta \\ &\quad \pm[(c_{11}-c_{44})^2\sin^4\theta + (c_{33}-c_{44})^2\cos^4\theta \\ &\quad + 2\sin^2\theta\cos^2\theta\{(c_{11}-c_{44})(c_{44}-c_{33}) \\ &\quad + 2(c_{13}+c_{44})^2\}]^{\frac{1}{2}}. \end{aligned} \quad (4)$$

The computations of the three wave velocities are clearly much simpler for the hexagonal metals than for the cubic metals. Figures 4(a)–(c) contain two sets of body wave curves based upon the independent c_ij

TABLE II.

c_ij data	Metal	θ_l *(a)	θ_l (b)	θ_{t_1} (a)	θ_{t_1} (b)	θ_{t_2} (a)	θ_{t_2} (b)
	Cd	3.13	2.94	1.41	1.34	1.76	1.68
	Mg	5.68	5.63	3.25	2.97	3.10	2.96
	Zn	4.05	4.03	2.03	2.07	2.71	2.68

* These letters refer to c_ij data employed as indicated in Figs. 4.

data of the indicated experimenters. By evaluating the areas under the various curves, one has a means for deriving the theoretical average velocities for the compressional and shear waves. Table II summarizes the findings.

ACKNOWLEDGMENT

The author is indebted to the Staff of the Computing Group of the Graduate Division of Applied Mathematics for assistance in the numerous calculations entailed.

Light Energy and Spectral Distribution for Some Important X-Ray Screens

HOLGER SKÖLDBORN
Department of Physics, Lund University, Sweden

(Received December 12, 1949)

The absolute energy of the light (ergs/cm² per *r*-unit) and its spectral distribution have been determined for four different types of x-ray fluorescent screens: Patterson *E*-2, Patterson *B*-2, Patterson *D*, and Neossal. The investigation was performed in accordance with an objective spectrophotometric process with the aid of a tungsten ribbon filament lamp as a standard, a light attenuator of reflection type, monochromator, multiplier photo-tube, and galvanometer.

The application of the method employed to the weak light intensities in question is made possible on account of the enormous light sensitivity of the multiplier photo-tube and the possibility of reproducible continuous variation of this sensitivity within very wide limits. The method has proved speedy and accurate and appears to be well suited for investigations of similar character, within the spectral range which may be measured by the commercial types of multiplier photo-tubes.

IN spite of the immense importance and varied applications which the x-ray fluorescent screens have developed in the last decades, one will search in vain for exact information concerning the absolute energy and spectral distribution of the emitted fluorescent light from different types of screens. The present investigation, which was planned early in 1947 after a proposition put forward by medical quarters, is intended to provide these desirable data for some of the most important commercial types of screens. We have chosen three of the firm du Pont's latest makes, Patterson *D*, Patterson *B*-2, and Patterson *E*-2, and the Neossal screen (of German make), which is probably still the most commonly used for fluoroscopy in Sweden. The physical properties of calcium tungstate x-ray screens have been treated in detail by Coltman, Ebbighausen, and Altar in an earlier article in this Journal.[1]

APPARATUS

The main problem of the investigation consisted in the determination of the spectral energy distribution of the fluorescent light from different x-ray fluorescent screens at a given x-ray radiation quality and a given measured *r*-intensity by comparison with the known spectral energy distribution from a temperature calibrated tungsten ribbon filament lamp. Integration of the spectral curves obtained in this way gives the absolute energy of the fluorescent light per *r*-unit incident on the screen.

When choosing the method of investigation we were guided by the following points of view. A practicable way for an investigation at these weak light intensities is, of course, the method of comparing in a microphotometer spectrograms recorded from the standard lamp, or the x-ray screen. To give reliable results, however, this method requires among other things an extremely elaborate technique for variable reduction of the filament brightness to a value closely approximating to that of the fluorescent screen at different wavelengths (on account of reciprocity failure of the photographic emulsion). RCA's latest extremely sensitive photo-cells of the multiplier type seemed to suggest the possibility of applying an objective spectrophotometric method, though the light output of the screen under the planned experimental conditions could not be expected to amount to more than 10–100 erg/cm²/sec. After a great number of preliminary experiments, comprising recording of spectrograms of several types of screens, testing of different photo-cells in connection with an appropriate monochromator etc., the final form of the apparatus was such that the method could be regarded as fully satisfactory for the present investigation. The apparatus employed will now be described in greater detail.

A tungsten ribbon filament lamp provided with a quartz window was used as standard source. The ribbon filament is about 10 mm long and 2 mm wide, and is supported in the bulb by heavy nickel leads. The current could easily be kept constant within 0.1 percent.

The lamp was temperature calibrated to the standard of the Physics Laboratory of the University of Lund with the aid of its optical pyrometer. At 13.00 amperes the brightness temperature outside the quartz window was determined at 2244°K for a point midway along the

[1] Coltman, Ebbighausen, and Altar, J. App. Phys. 18, 530–544 (1947).

filament. This current was used all through the investigation.

When computing the spectral distribution of the radiation emitted by the ribbon filament lamp one must know the true temperature of the filament inside the bulb, the spectral emissivity of tungsten, and the spectral transmission of the quartz window within the intended working range of wave-lengths (4000–7000A). For the spectral emissivity the values determined by Hamaker[2] were used. For the spectral transmission of the quartz window the value 0.92 was used within the whole range of wave-length in question. The error possibly introduced in this way (the lamp was an unused specimen) is, according to Hamaker's results from measurements of absorption in quartz windows of different conditions,[2] not likely to exceed 0.5 percent anywhere. The conversion from the brightness temperature to the true temperature was accomplished by the following formula, derived from Planck's radiation formula:

$$\frac{1}{T_t}=\frac{1}{T_{BO}}+\frac{\lambda}{0.4343C_2}\log_{10}(\epsilon_\lambda\cdot t_\lambda), \tag{1}$$

where T_t = true temperature of the filament, T_{BO} = brightness temperature of the filament outside the window as determined by the optical pyrometer, λ = mean effective wave-length in cm of the red filter used in the optical pyrometer, C_2 = the second constant in Planck's radiation formula, ϵ_λ = the spectral emissivity of tungsten at λ, and t_λ = the spectral transmission of the quartz window at λ.

With $T_{BO}=2244$°K, $\lambda=6650\times10^{-8}$ cm, $C_2=1.435$, $\epsilon_\lambda=0.430$, $t_\lambda=0.920$, the true temperature of the filament is 2483°K.

The spectral energy distribution outside the lamp was obtained by substitution of the above value of the true temperature in Planck's radiation formula, and correcting the result for the spectral emissivities of tungsten and for the spectral transmission of the quartz window:

$$B_\lambda d\lambda=\frac{C_1\lambda^{-5}}{e^{C_2/\lambda T}-1}\epsilon_\lambda t_\lambda d\lambda. \tag{2}$$

In the above formula $B_\lambda d\lambda$ denotes the brightness in ergs/cm²/second per unit solid angle within the wave-length interval $d\lambda$ (λ in cm and C_1 its corresponding value).

An area of 5 mm×0.3 mm of the central portion of the filament was used for projection on the slit of the monochromator. At the pyrometric calibration no temperature differences were observed within this central portion. From Planck's radiation law one derives the percentage change in emitted energy at a given wave-length for a change in true temperature to be

$$\left(\frac{dE_\lambda}{E_\lambda}\right)\text{percent}=100\left(\frac{C_2}{\lambda T^2}\right)dT. \tag{3}$$

For a true temperature of 2483°K the percentage change of E_λ per degree is 0.58 percent at 4000A and 0.36 percent at 6500A. The error in the true, pyrometrically determined temperature is estimated to ±3°K. This means an error in E_λ of 1.7 percent at 4000A and 1.1 percent at 6500A. Considering all sources of error it appears that the combined error in the computed energy varies from about 4 percent at 4000A to about 3 percent at 6500A with an average for the middle of the spectrum of about 3.5 percent.

To make possible a spectrophotometric comparison (in the wave-length range 4000–7000A) between the brightness of the standard lamp filament and the fluorescent screen, the brightness of the former had to be reduced by a factor, which turned out to be of the order of 10^{-5}. The sensitivity of the type of multiplier photo-tube used being variable continuously and reproducibly within very wide limits, this reduction could be obtained by a suitable non-variable light "attenuator" and further reduction arrangements avoided.

A light attenuation of the order of magnitude desired can be accomplished by successive reflections from four air-glass surfaces, and an apparatus designed and constructed according to this principle is described in an article cited above.[1] We constructed such an attenuator with great care of four crown glass prisms, arranged as in Fig. 1, inside a brass box. The principle is immediately clear from the figure; concerning the details we will be as brief as possible, in our reference to the article mentioned. All the prism surfaces, except the ones used for reflection, were covered with a highly absorbent coating of carbon black silted up in zapon lacquer. The non-reflecting and absorbing coating obtained serves to remove to a very high degree the light transmitted into the prism.

According to Fresnel's formula one obtains for the light polarized in the plane of incidence after four successive reflections the following expression for the

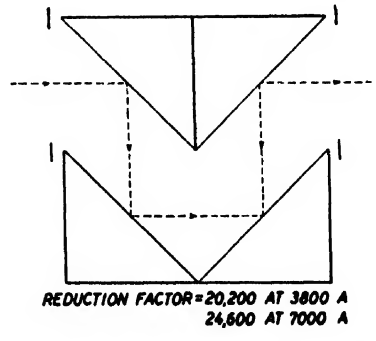


FIG. 1. Schematic arrangement of reflection-type attenuator.

[2] H. C. Hamaker, "Reflectivity and Emissivity of Tungsten," Diss. Utrecht (1934).

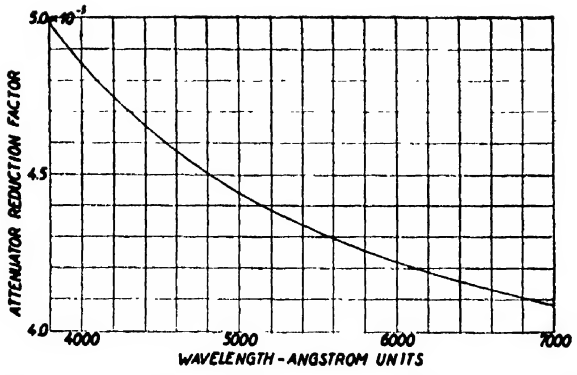


FIG. 2. Reduction factor of attenuator as a function of wave-length

intensity:

$$R(I)=\tfrac{1}{2}\left[\frac{\sin(i-i')}{\sin(i+i')}\right]^{8}. \tag{4}$$

The polarization being practically complete, the intensity of the light polarized at right angles to the plane of incidence can be neglected (it turns out to be about 9000 times fainter than the other component).

With i fixed at 45° the above formula may be written:

$$R(I)=\tfrac{1}{2}[\tan(45-i')]^{8}, \tag{5}$$

where i' depends upon n (Snell's law), which in its turn varies with the wave-length. R is thus itself a function of the wave-length, on account of which it was necessary to determine the dispersion curve of the prism material for the range of spectrum to be studied. For this purpose n was determined for one of the prisms for each of the three mercury lines 4046.6, 4358.3, and 5460.7A. The resulting indices were, respectively 1.5302, 1.5367, and 1.5187. This determination was carried out with a Gaertner spectrometer type 568; each value had a standard deviation of 1 unit in the fourth decimal place. With the values of the wave-length and their respective indices, stated above, the constants of the Hartmann interpolation formula were obtained:

$$n=1.4977+\frac{83.969}{\lambda-1462.9}. \tag{6}$$

With intervals of 100A n was determined from (6) for the range 3800–7000A, after which i' from Snell's law and finally the reduction factors from (5) were obtained as a function of the wave-length. The last-mentioned relation is represented by the curve in Fig. 2. The curve in Fig. 3, obtained by applying these reduction factors to the spectral energy distribution curve of the filament outside the lamp, shows the resulting spectral brightness of the standard lamp filament as seen through the attenuator.

As mentioned above the light from the attenuator is practically completely polarized, while the light from the fluorescent screen is unpolarized. To avoid errors which may arise, if the light loss in the monochromator is not the same for both sources, the attenuator was mounted so that the plane of the normals of the reflecting surfaces made an angle of 45° with the slit. In this way the light received by the monochromator may be regarded as isotropic.

The errors in the computed reduction factors are evaluated as follows:

1. An error of 1 unit in the fourth place of $n-1$ results in an error in R of less than 0.1 percent.
2. The error in R to be expected in consequence of the dispersion curve being calculated from only three index determinations is estimated to at most 0.1 percent for the range 4000–7000A.
3. The beam through the attenuator, as defined by a rectangular stop on the collimator lens and the focal distance of this lens, had an angular size of about 5°×3°. Since R was determined for a constant angle of incidence of 45°, it was necessary to obtain an estimate of the error introduced by assuming $i=45°$ for the whole beam. Calculation shows that the error, which is negative, amounts to about 0.5 percent. When this deviation in the reduction factors has been corrected of, any possibly remaining error is unlikely to exceed 0.2 percent.
4. The construction as well as the alignment of the attenuator was performed with great care so that the resulting error in R due to want of accuracy is probably less than 0.5 percent.

From these points we may conclude that the attenuator functions with an accuracy of 1 percent.

The dispersing instrument used was a "Hilger Constant Deviation Wave-Length Spectrometer," type D 187. It has $f=285$ mm for both lenses, and an angular aperture of 1:9. An additional rectangular stop of the size 25 mm×14 mm was fitted close to the collimator lens. Both slits were adjusted to a width of 0.3 mm and a height of 5 mm. The reliability of the wave-length scale was tested with the aid of a line spectrum. After a careful adjustment of the prism it appeared that the wave-length readings for the range of 4000–6500A showed a maximum error of about ±1A. Thus the accuracy is entirely satisfactory for our purpose.

The experimental arrangement is shown in Fig. 4. By means of a mirror and a condensing lens, either the filament or the screen sample could be projected on to the collimator slit at a magnification of 1:1. The mirror (an aluminized plane mirror) could be rotated 90° and mounted in two accurately fixed positions. The alignment was checked by illuminating the emergent slit of the monochromator by a strong light source, the slit then being imaged at the center of the filament.

The photo-cell was a RCA type 1P21 multiplier photo-tube. The light sensitive surface is made up of a cesium-antimony layer on a nickel base and the spectral response curve is of the S-4 type. A number of photocells of the multiplier type often show great, individual differences with respect to sensitivity and dark current. The cell employed was a selected specimen with a high value of the signal-to-noise ratio. We considered a minimizing of the dark current by reducing the temperature,[3] but this was found to be unnecessary for the

[3] R. W. Engstrom, J. Opt. Soc. Am. 37, 420 (1947).

present investigation. The photo-cell was placed in a suitable light-proof brass holder that was screwed on to the site of the eyepiece of the monochromator. The distance from the slit to the photo-cathode was about 28 mm.

The dynode voltage of the photo-cell was provided from a voltage-stabilized mains-connected supply circuit, designed and constructed by Stigmark[4] at the Physics Laboratory. The potential dividers for the dynodes consisted of 9 exactly equal 50,000 ohms wire-wound resistors. The current through these resistors could be continuously varied from 1 to 2 milliamperes and read on a milliammeter, 2 milliamperes corresponding to a voltage drop of 100 volts per stage. When checking the properties of the voltage supply it was found that a change of the mains voltage (220 v) of ± 10 v caused a change of the total dynode voltage (1000 v) of 0.05 percent, corresponding to a variation of the anode output current of about 0.5 percent. This accuracy was considered to be quite satisfactory.

The anode output current was measured by a "Multiflex" mirror galvanometer with a sensitivity of $5 \cdot 10^{-10}$ ampere/mm. The length of the scale being 200 mm, full deflection corresponded to 0.1 microampere.

After the voltage supply had been tested and found to be sufficiently constant, it remained to investigate the fatigue effects of the photo-cell and the linearity of response of the photo-cell-galvanometer combination. Considering what has been written about the fatigue of multiplier photo-tubes,[3, 5] we decided from the beginning not to use higher anode current than 0.1 microampere, especially as this could be done without actual inconvenience during the present investigation. From repeated tests made at constant illumination corresponding to an anode current of the order of magnitude mentioned, it appeared that the fatigue effect must be given due attention even at this low light intensity. The galvanometer deflection decreased 3 to 5 percent during the first half-hour; thereafter, however, the fatigue proceeded extremely slowly. At the performance of the investigation we took this fact into account by fatiguing the photo-cell at 0.1 microampere for about one hour before each series of observations. Moreover the series of observations were carried out according to a scheme which ensured that a continued, slight fatigue effect could have practically no influence on the result.

The proportionality between the intensity of the incident light and the deflection of the galvanometer was checked according to a method given by Hamaker,[2] the principle of which will be clear from Fig. 5. By means of the condensing lens L two light sources LS and LS' were focussed on the slit of the monochromator, LS directly and LS' by way of the reflecting glass plate G; S and S' are two shutters. The check simply consisted of observing whether the sum of the deflections caused by the separate intensities was equal to the deflection caused by the sum of the intensities. The advantages of this method are evident; no changes in the experimental arrangements are needed, the stopping down of the light beam remains unaltered and all we have to do is to use the shutters and vary the electric current through the light sources LS and LS'.

By this method the proportionality of deflection and intensity was checked for 10 different positions on the galvanometer scale. Three such series of observations were carried out for three different wave-length settings of the monochromator (4500, 5400, and 6300A). From these checks it was clear that the deflection of the galvanometer varied proportionally with the intensity of the incident light within 0.5 percent for deflections between 20 and 200 mm (full deflection). The small deviations discovered most certainly originate from the galvanometer employed; concerning the photo-multiplier tube there is no reason to doubt a strict proportionality between light intensity and anode current until above 100 microamperes.

At a distance of 35 cm from the focus of the air-cooled x-ray tube there was a lead stop with a hole 5 cm square. Behind this stop an additional filter of 1 mm Al was applied, and in front of the stop, at a distance of 37.5 cm from the focus of the tube, the screen sample was placed in a holder (Fig. 4). The x-ray intensity was measured at the exact place of the screen by means of an ionization chamber (so-called secondary standard), which, for the radiation quality employed, had been specially calibrated to the primary standard of the Radiophysical Institute in Stockholm.

EXPERIMENTAL PROCEDURE

Before the light intensity measurement could suitably commence, there remained the determination of the relative sensitivity of the photo-cell as a function of the applied dynode voltage. Preliminary experiments showed

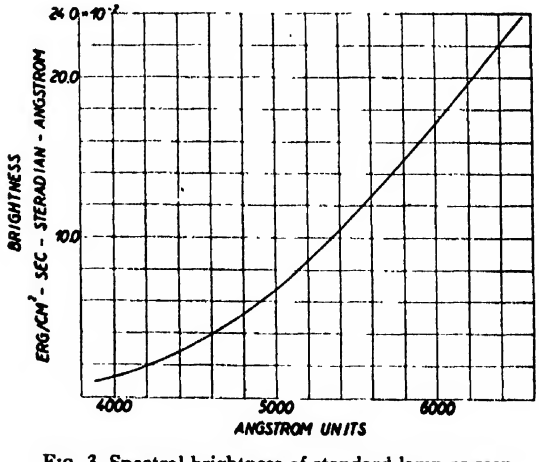


FIG. 3. Spectral brightness of standard lamp as seen through attenuator.

[4] K. Lidén, and L. Stigmark, "Some Experiences from Investigations of the Spectral Intensity Distribution of Weak Light Sources by Means of Multiplier Photo-Tubes," (Manuscript 1948).

[5] K. G. Kessler and R. A. Wolfe, J. Opt. Soc. Am. 37, 133 (1947).

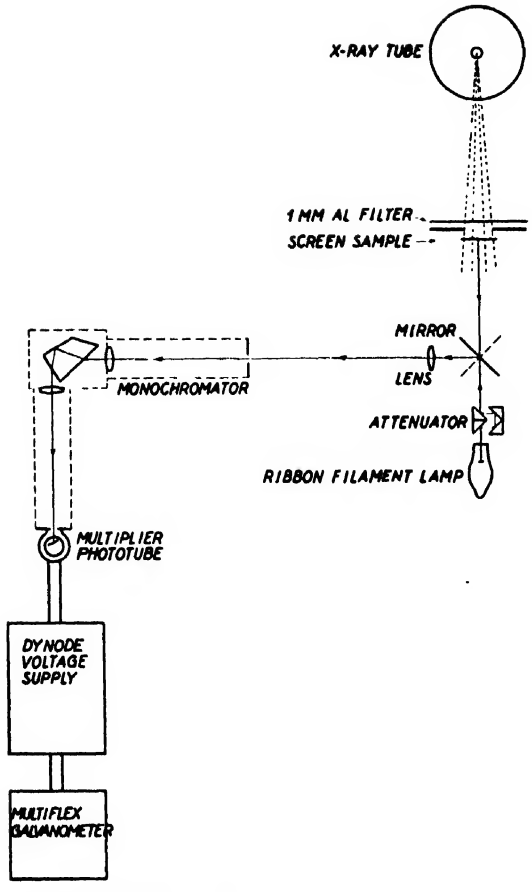


FIG. 4. Experimental arrangement.

that to obtain sufficient galvanometer deflection from both the screen sample and the standard lamp (seen through the attenuator) within the wave-length range to be studied, a variation of the dynode voltage from about 50 to 90 volts per stage was required. The upper limit was determined by the fact, that the signal-to-noise ratio of the photo-cell employed was rather unfavorable at higher voltages. As the accuracy and reproducibility of the sensivity curve of the photo-cell was of decisive importance for our method of measuring, the determination of this curve was given special attention. Sensitivity curves were determined for five different photo-multiplier tubes, and their reproducibility was investigated with respect to the fatigue effects, etc. on several occasions. It was evident from these experiments that the relative sensitivity as a function of the dynode voltage was independent of the fatigue effects and reproducible within limits of error exclusively determined by the accuracy of observation of the instruments.

When determining these sensitivity curves the following method was employed, which, among other things had the advantage of being applicable without change of the experimental arrangements. The electric current through the standard lamp filament was regulated in such a way that the galvanometer showed full deflection at a dynode voltage of 90 volts, after which the deflection was observed for successively lower dynode voltages, as far as the scale of the galvanometer permitted. Then the intensity of the lamp was increased, until full deflection was achieved at a lower dynode voltage, and so forth. Three different light intensities permitted a variation of the dynode voltage from 90 to 50 volts; in this way the curve was joined at two points, the factor of the joint being obtained as the average of two or three double readings. The sensitivity curve of the photo-cell employed (Fig. 6) was computed from the average of four such series of readings. The relative amplification factor varies from 100.0 to 0.910 at a variation of the dynode voltage from 90 to 50 volts. The error in the ratio of two arbitrary values of the curve was calculated to be at most 2 percent.

The determination of the spectral energy distribution of the fluorescent screens was now performed in the following way. After the photo-cell had been fatigued for about an hour and the standard lamp had been allowed to assume temperature equilibrium, the readings for two complete series were performed according to the scheme: $LSL'S'SLS'L'$. The readings were taken with intervals of 100A for the lamp as well as the screen. L and S denote readings of the deflections for the intensities of the standard lamp and the fluorescent screen, respectively, starting at the violet end of the spectrum; L' and S' readings for the same wave-length starting at the red end of the spectrum. This scheme was used in order to eliminate completely the fatigue effect on the average of each determination. To read the deflection for the standard lamp as well as for the screen point by point turned out to be unpractical, as in such a case the dynode voltage had to be readjusted very often. With the above method several readings often could be made in succession before readjusting the dynode voltage to a new value.

In connection with each series of readings three measurements of the x-ray intensity were carried out at the position of the screen according to the scheme $RLSRL'S'R$, where R denotes an r-measurement. The x-ray intensity could be kept constant within 1 percent (temporary fluctuations of the tube current). The accuracy of a separate r-intensity measurement was about 1 percent. The measured x-ray intensity was on an average 0.901 r-unit per second.

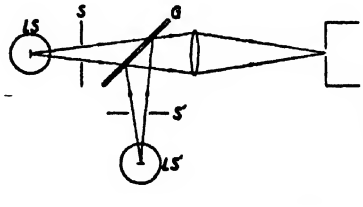


FIG. 5. Arrangement for check of proportionality between light intensity and galvanometer deflection.

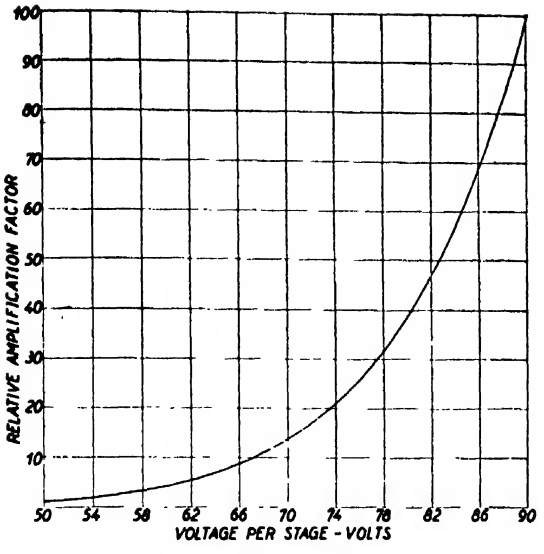


FIG. 6. Variation of relative amplification factor with applied voltage per stage for 1P21 tube used.

EXPERIMENTAL RESULTS

The spectral energy distribution of the screens investigated is plotted in Fig. 7. The spectral curves are given in terms of light energy emerging from the screen per r-unit incident on the screen. All results refer to the same x-ray quality, *viz.* 80 kv constant potential x-rays, filtered by one mm Al (tube current 9 mA). The filter equivalent of the wall of the x-ray tube was about 1.0 mm Al. The experimentally determined points succeed each other so smoothly that every point, if plotted, would be found on the drawn lines in Fig. 7.

The values of the absolute energy of the emerging fluorescent light obtained by numerical integration of the curves in Fig. 7 are given in Table I. The second column of this table gives the measured brightness of the screens in ergs/cm²/unit solid angle/r-unit incident on the screen. This applies to light emitted in the direction of the normal. The values of the total light emission in the third column have been calculated on the assumption that the departure from Lambert's law is represented by a factor of 0.92. This value was found by Coltman, Ebbighausen, and Altar[1] for a $CaWO_4$ screen excited by 80 kv constant potential x-rays, filtered by 0.5 mm Cu and 1 mm Al. In the present case, however, this value may be inaccurate with an error of one or two percent, owing to variation with radiation quality and material of the screen.

Of the screen samples investigated Patterson *D*, Patterson *B*-2, and Patterson *E*-2, were fresh samples obtained directly from the manufacturer (the Patterson Screen Division of E. I. du Pont de Nemours & Company, at Towanda, Pennsylvania). The Neossal screen, on the other hand, was a several years old sample, which, however, was unused and kept light-proof. All the screens were of the ordinary type with the phosphor layer applied to a white mounting card. The phosphor layer of the Neossal and the Patterson type *E*-2 screens were covered with a protective film of lacquer. On the Patterson type *D* screen, which fairly recently was developed for photo-fluorography, the fluorescent layer is composed of silver-activated zinc sulfide; the other three, which are mostly used for fluoroscopy, may be classed as zinc-cadmium sulfide screens. The Patterson type *B*-2, which was commercially available early last

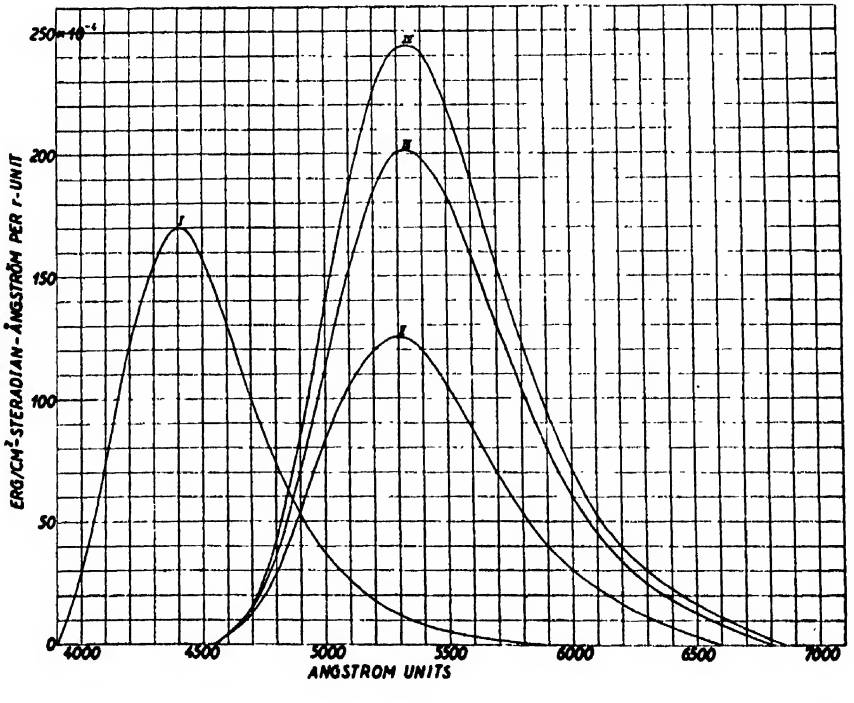


FIG. 7. Fluorescent spectra of four important x-ray screens, excited by 80 kv constant potential x-rays, filtered by 1 mm aluminium. I: Patterson type *D*. II: Neossal. III: Patterson type *B*-2. IV: Patterson type *E*-2.

TABLE I. Brightness and total emission of light of four x-ray fluorescent screens (calculated from Fig. 7).

Screen	Brightness ergs/cm²/steradian/r-unit	Total light emitted from screen ergs/cm²/r-unit
Patterson type *D*	11.8	34.0
Neossal	11.0	31.8
Patterson type *B*-2	18.4	53.2
Patterson type *E*-2	22.1	64.0

year was stated to be 40 percent brighter than the next preceding type Patterson *B*, a statement that we have had the opportunity of verifying. We have also included the results for the Patterson type *E*-2, which is stated to be of the same character as the type *B*-2, but with a protective cleanable surface, owing to the fact that it was found to be not less than 20 percent brighter than the type *B*-2 of the samples received at the same time. The Patterson type *D* with unprotected surface is also manufactured with a protective cleanable surface, and then called type *D-C*. These two variants turned out to be of equal brightness.

The accuracy of an investigation of this kind depends on several factors. If the contributions from all sources of error are summed up according to the calculations and estimates imparted at the description of the corresponding part of the apparatus, and of the experimental procedure, it appears that the results stated above will be correct with a maximum error of about 9 percent.

Unbalanced Terminations on a Shielded-Pair Line*

K. TOMIYASU
Division of the Sperry Corporation, Great Neck, New York
(Received December 7, 1949)

Theoretical calculations indicate that the unbalanced component of current on a two-wire or shielded-pair transmission line can be determined by comparing the different standing-wave distributions on the line conductors. The measurement of reflected components from an unbalanced radiative termination on a slotted shielded-pair line is described. The existence of a condition of minimum unbalanced reflected component from end-coupled antennas is shown experimentally.

INTRODUCTION

SHIELDED-PAIR and two-wire transmission lines differ from coaxial lines in that the former can support simultaneously two transverse electromagnetic (*TEM*) propagating modes. One of the modes called the balanced mode is the conventional transmission line mode where the currents on the two conductors are *opposite* in phase and equal in magnitude at all cross sections. For the other mode called the unbalanced mode, the currents are *in* phase and equal in magnitude and hence the scalar potential difference between the conductors is zero at all cross sections. The unbalanced mode radiates energy from an open two-wire line but does not radiate from a shielded-pair line. Whether the waves are standing or progressive the usual physical effect of the coexistence of the balanced and unbalanced modes is that the currents on the two conductors are unequal in magnitude and of arbitrary phase at all cross sections. This further indicates that the standing-wave distributions will be different on the two inner conductors if a shielded-pair line is excited by a balanced mode and terminated in an unbalanced structure. The reflected balanced and unbalanced modes can be measured from this difference. Measurements have been made on a slotted shielded-pair line which was terminated in unbalanced radiative terminations. Some of the representative measurements are included in this paper.

INCIDENT AND REFLECTED WAVES

The difference between the standing-wave distributions on the two conductors which yields the total characteristics of an unbalanced termination is dependent on the relative phases and amplitudes of four modes or waves: balanced and unbalanced, incident and reflected modes. Since these modes can be added as well as decomposed readily the complex voltage components on the conductors can be written as follows.

$$V_{1BI} = -V_{2B1} = A_0 \exp(-\gamma_B X + j\omega t), \quad (1)$$

$$V_{1BR} = -V_{2BR} = B_0 \exp(\gamma_B X + j\omega t), \quad (2)$$

$$V_{1UR} = V_{2UR} = C_0 \exp(\gamma_U X + j\omega t), \quad (3)$$

$$V_{1UI} = V_{2UI} = D_0 \exp(-\gamma_U X + j\omega t), \quad (4)$$

where B stands for balanced; U stands for unbalanced; I stands for incident; R stands for reflected; A_0, B_0, C_0, D_0 are complex constants; $\gamma = \alpha + j\beta$ is the complex propagation constant; $+X$ is the direction away from generator. γ_B may differ from γ_U for dielectric-filled shielded-pair lines as well as open two-wire lines. The time factor $e^{j\omega t}$ may be omitted for convenience. The

* The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (ONR), the Signal Corps of the U. S. Army, and the U. S. Air Force, under ONR Contract N5-ori-76, T. O. 1.

resultant voltages on each line are

$$V_1 = V_{1BI} + V_{1BR} + V_{1UR} + V_{1UI} = V_{1B} + V_{1U}, \quad (5)$$

$$V_2 = V_{2BI} + V_{2BR} + V_{2UR} + V_{2UI} = V_{2B} + V_{2U}. \quad (6)$$

The addition of these voltage components is illustrated in Fig. 1. Since $V_{1B} = -V_{2B} = 1/2(V_1 - V_2)$ and $V_{1U} = V_{2U} = 1/2(V_1 + V_2)$ as shown in Fig. 1, the balanced and unbalanced components can be determined provided the amplitudes and phase difference of V_1 and V_2 are known.[1] The standing-wave distribution curves which can be obtained from $|V_1|$ and $|V_2|$ of Eqs. (5) and (6) are symmetrical within each half-wave-length for all values of the complex constants.

The desired reflection coefficients which characterize the unbalanced termination can be obtained by considering each line separately. For the general case of four modes, the reflection coefficients can be defined for the terminations as

$$\Gamma_1 = \left|\frac{V_{1BR} + V_{1UR}}{V_{1BI} + V_{1UI}}\right| \underline{/\phi_1'} = \left|\frac{V_{1R}}{V_{1I}}\right| \underline{/\phi_1'}, \quad (7)$$

$$\Gamma_2 = \left|\frac{V_{2BR} + V_{2UR}}{V_{2BI} + V_{2UI}}\right| \underline{/\phi_2'} = \left|\frac{V_{2R}}{V_{2I}}\right| \underline{/\phi_2'}. \quad (8)$$

For this case of balanced and unbalanced, incident and reflected modes on a line where $\gamma_B = \gamma_U$, six unknowns must be determined for the complex constants, A_0, B_0, C_0, and D_0. Since no additional information on an unbalanced termination can be obtained by having an unbalanced *incident* mode, it is mathematically and experimentally convenient to require D_0 to be equal to zero. This reduces the number of unknowns to four, and these can be obtained from the standing-wave ratios and positions of voltage minima of the standing-wave distributions. The other two unknowns necessary for the general case may be the time-phase difference[2] and relative magnitudes of the distribution extremes.

When $D_0 = 0$, the reflection coefficients at the load after arbitrarily assigning a zero phase angle for A_0, are

$$\Gamma_1 \equiv \frac{B_1}{A_0} = \frac{B_0 + C_0}{A_0} = \frac{|B_0 + C_0|}{A_0} \underline{/\phi_1}, \quad (9)$$

$$\Gamma_2 \equiv \frac{B_2}{A_0} = \frac{B_0 - C_0}{A_0} = \frac{|B_0 - C_0|}{A_0} \underline{/\phi_2}, \quad (10)$$

where

$\phi = 2\beta X_{\min} \pm \pi$,

$\beta = 2\pi/\lambda$ = phase constant,

$X_{\min}$ = distance from the load to a voltage minimum

[1] See, for example, R. King, "Coupled antennas and transmission lines," Proc. I.R.E. 31, 626–640 (November, 1943).

[2] For measurement of time-phase differences, see, for example, T. Morita, Cruft Laboratory Technical Report No. 66, Harvard University, Cambridge, Massachusetts.

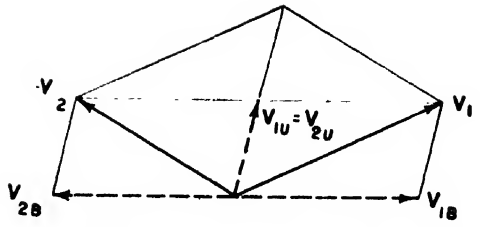


FIG. 1. The components of two unequal voltages.

and the voltage standing-wave ratios are

$$VSWR_1 = \frac{1 + |\Gamma_1|}{1 - |\Gamma_1|}, \quad (11)$$

$$VSWR_2 = \frac{1 + |\Gamma_2|}{1 - |\Gamma_2|}. \quad (12)$$

Defining

$$V_I \equiv A_0, \quad (13)$$

$$V_{BR} \equiv \frac{B_1 + B_2}{2} = B_0, \quad (14)$$

$$V_{UR} \equiv \frac{B_1 - B_2}{2} = C_0 \quad (15)$$

the following equations may be written

$$\frac{V_{BR}}{V_I} = \frac{\Gamma_1}{2} + \frac{\Gamma_2}{2} = \Gamma_1\left(\frac{1}{2} + \frac{1}{2}\frac{\Gamma_2}{\Gamma_1}\right) = \Gamma_1 F \equiv \Gamma_B, \quad (16)$$

$$\frac{V_{UR}}{V_I} = \frac{\Gamma_1}{2} - \frac{\Gamma_2}{2} = \Gamma_1\left(\frac{1}{2} - \frac{1}{2}\frac{\Gamma_2}{\Gamma_1}\right) = \Gamma_1 G \equiv \Gamma_U, \quad (17)$$

where

$$F \equiv \frac{1}{2} + \frac{1}{2}\frac{\Gamma_2}{\Gamma_1}, \quad (18)$$

$$G \equiv \frac{1}{2} - \frac{1}{2}\frac{\Gamma_2}{\Gamma_1}. \quad (19)$$

The quantity Γ_U is actually a "transfer" reflection coefficient since the usual connotation implies that the reflected mode is the same as the incident. The complex parameters F and G can be redefined as:

$$F = \tfrac{1}{2} + \tfrac{1}{2}M\underline{/\alpha} = |F|\underline{/\theta_F}, \quad (20)$$

$$G = \tfrac{1}{2} - \tfrac{1}{2}M\underline{/\alpha} = |G|\underline{/\theta_G}, \quad (21)$$

where

$$M = \frac{|\Gamma_2|}{|\Gamma_1|},$$

$$\alpha = \phi_2 - \phi_1,$$

$$|F| = +\tfrac{1}{2}(1 + M^2 + 2M\cos\alpha)^{\frac{1}{2}}, \quad (22)$$

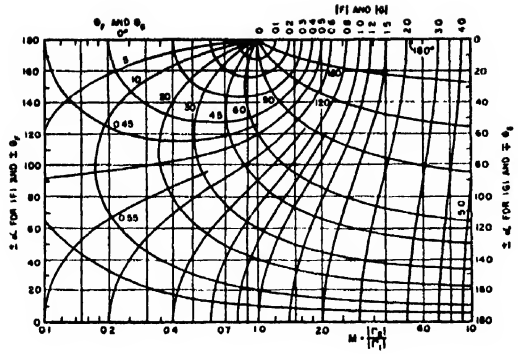


FIG. 2. Parametric plots of $|F|$, $|G|$, θ_F, and θ_G as functions of α and M.

$$|G| = +\tfrac{1}{2}(1+M^2-2M\cos\alpha)^{\frac{1}{2}}, \tag{23}$$

$$\theta_F = \tan^{-1}\frac{M\sin\alpha}{1+M\cos\alpha}, \tag{24}$$

$$\theta_G = \tan^{-1}\frac{-M\sin\alpha}{1-M\cos\alpha}. \tag{25}$$

Orthogonal parametric plots of $|F|$, $|G|$, θ_F and θ_G which are functions of M and α are shown in Fig. 2. Using these plots and Eqs. (16) and (17), the balanced and unbalanced reflected modes can be determined from the measurable quantities $|\Gamma_1|$, $|\Gamma_2|$, ϕ_1, and ϕ_2. It is evident from the plots that there is only one condition for a completely balanced termination: $|\Gamma_1| = |\Gamma_2|$ and $\phi_1 = \phi_2$ simultaneously ($M=1$, $\alpha=0$).

EQUIVALENT TERMINATING IMPEDANCES

From Eqs. (9)–(10) and (16)–(17) it can be seen that the reflections from an unbalanced termination may be represented by either Γ_1 and Γ_2 or Γ_B and Γ_U. Since the termination can be completely represented by two independent impedances related to Γ_1 and Γ_2, a third impedance as required for example in a delta-network is unnecessary. When the generator supplies power in only the balanced mode, Γ_B is a measure of all losses due to both balanced and unbalanced modes. It may be considered incidental if the unbalanced reflected mode propagates along the input line.

COUPLING OF VOLTAGE PROBE TO OPPOSITE LINE

When used on a shielded-pair or two-wire line, a voltage probe will respond to electric fields due to the charges on both conductors. If the coupling to the remote conductor is significant the probe will not yield the true standing-wave distribution on the near conductor. The true reflection coefficients can be determined from the measured reflection coefficients if the "opposite-conductor" coupling is known.

Assuming that the two voltage probes are equally coupled to the shielded-pair line, the effective voltages at the probes are

$$V_1' = A_0(1-K_1)e^{-i\beta X} + B_0(1-K_1)e^{i\beta X} + C_0(1+K_2)e^{i\beta X}, \tag{26}$$

$$V_2' = -A_0(1-K_1)e^{-i\beta X} - B_0(1-K_1)e^{i\beta X} + C_0(1+K_2)e^{i\beta X}, \tag{27}$$

where K_1 = fractional *decrease* caused by coupling of charges of opposite polarity and K_2 = fractional *increase* caused by coupling of charges of same polarity. The modified and measurable reflection coefficients are

$$\Gamma_1' = \frac{B_0(1-K_1)+C_0(1+K_2)}{A_0(1-K_1)} = \Gamma_B + \frac{1}{K_0}\Gamma_U, \tag{28}$$

$$\Gamma_2' = \frac{B_0(1-K_1)-C_0(1+K_2)}{A_0(1-K_1)} = \Gamma_B - \frac{1}{K_0}\Gamma_U, \tag{29}$$

where

$$K_0 \equiv \frac{1-K_1}{1+K_2}. \tag{30}$$

Therefore, the true reflection coefficients are

$$\Gamma_B = \frac{\Gamma_1'+\Gamma_2'}{2} = \Gamma_1'\left(\frac{1}{2}+\frac{1}{2}\frac{\Gamma_2'}{\Gamma_1'}\right), \tag{31}$$

$$\Gamma_U = K_0\frac{\Gamma_1'-\Gamma_2'}{2} = K_0\Gamma_1'\left(\frac{1}{2}-\frac{1}{2}\frac{\Gamma_2'}{\Gamma_1'}\right). \tag{32}$$

Except for the factor K_0 and the primes, the equations are identical to the previous coefficients given in (16)

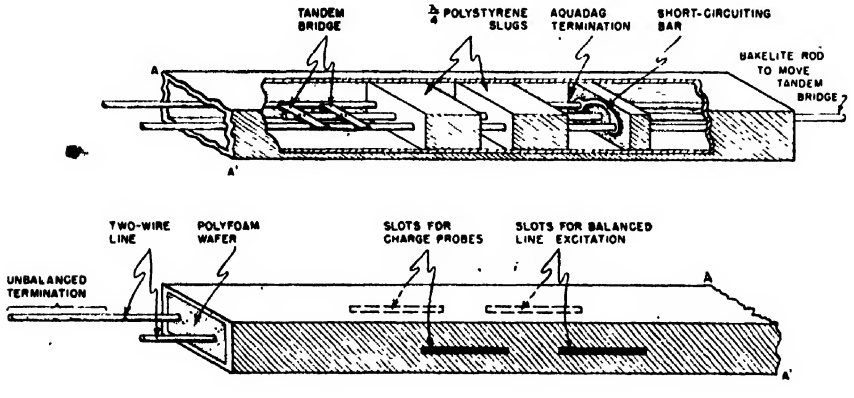


FIG. 3. Slotted shielded-pair transmission line with open end.

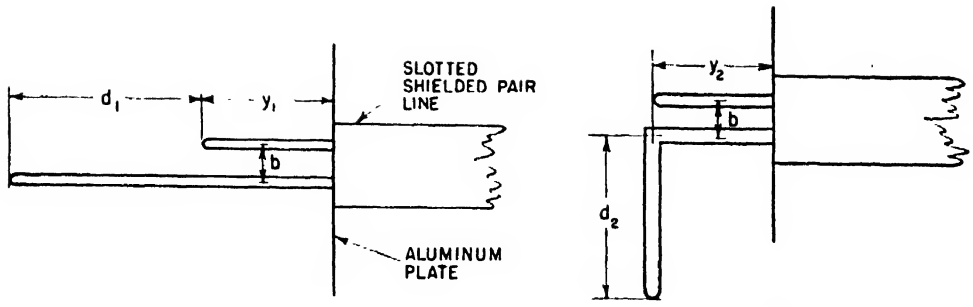


FIG. 4. (a) Collinear end-coupled antenna. (b) Perpendicular end-coupled antenna.

and (17). The quantity K_u can be determined experimentally. Moreover, in the presence of opposite-conductor coupling, the distribution curves are symmetrical within each half-wave-length.

RATIO OF VOLTAGE MAXIMA

Occasionally the *VSWR* on one conductor may be exceedingly large so that the minimum deflection is masked by the inherent noise of the detector system. For the case of zero incident unbalanced wave, the reflection coefficient on one line is related to both the reflection coefficient of the other line and the ratio of the two voltage maxima. This relationship is written as

$$\frac{|V_{2\,\max}|}{|V_{1\,\max}|}=\frac{|A_0|+|B_2|}{|A_0|+|B_1|}=\frac{1+|\Gamma_2|}{1+|\Gamma_1|}. \tag{33}$$

For a square-law detector, the ratio can be expressed in units of decibel as

$$S_m=40\log_{10}\frac{1+|\Gamma_2|}{1+|\Gamma_1|}\text{db}. \tag{34}$$

For convenience, Eq. (34) can be plotted for parametric values of $|\Gamma_2|$ or $|\Gamma_1|$. The maximum range of S_m is ± 12 db. If the measured S_m, $|\Gamma_1|$ and $|\Gamma_2|$ do not satisfy Eq. (34) there exist *both* incident and reflected unbalanced modes.

THE SLOTTED SHIELDED-PAIR LINE

The reflection coefficients of unbalanced terminations were measured on a slotted shielded-pair line. A cutaway view of the measuring line is shown in Fig. 3. The shielded-pair line consists of a standard $\frac{1}{2}$ in.$\times$1 in. *X* band copper wave guide and $\frac{1}{8}$ in. O.D. brass inner conductors spaced 0.495 in. between centers. An unbalance squelcher[3] in the generator circuit gave assurance that the balanced mode would be predominantly excited on the line. The 750 mc/sec. r-f signal was amplitude modulated at 1050 c.p.s. With an *SPDT* coaxial switch in the dual charge-probe circuit, standing-wave distributions could be measured on either conductor.[4] From an unbalanced termination the reflected balanced mode does not proceed beyond the tandem bridge, however, the reflected unbalanced mode is completely absorbed at the matched load provided by the Aquadag termination and two quarter-wave-length polystyrene slugs.[5]

With the inner conductors supported only by Polyfoam wafers, the characteristic impedances of the two modes of transmission were determined by measuring capacitances at an audiofrequency.[1] The measured characteristic impedances of the air-dielectric line for the balanced mode was $R_{cB}=165$ ohms and for the unbalanced mode $R_{cU}=39.3$ ohms. The opposite-conductor coupling was also determined under the same conditions in the following way. The line conductor to which the probe is coupled is connected to one output terminal of the audio oscillator. The second line conductor can be connected to either of two potentials: (a) to the other

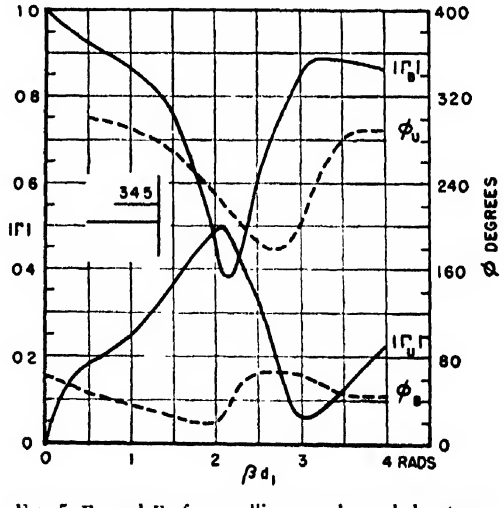


FIG. 5. Γ_B and Γ_U for a collinear end-coupled antenna when $y_1=34.5$ cm.

[3] K. Tomiyasu, Rev. Sci. Inst. 19, 675–677 (1948).

[4] K. Tomiyasu, Cruft Laboratory Technical Report No. 86, Harvard University (September 6, 1949).

[5] K. Tomiyasu, "Double-slug impedance matching," Cruft Laboratory Technical Memorandum No. 6, Harvard University (August 1, 1949)

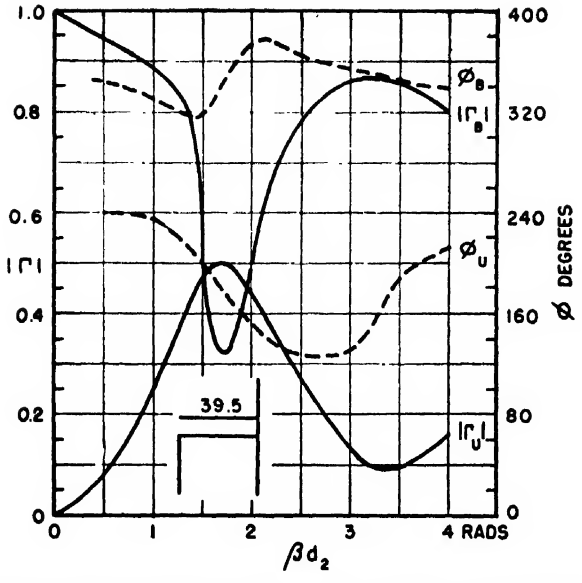


FIG. 6. Γ_B and Γ_U for a perpendicular end-coupled antenna when $y_2=39.5$ cm.

output terminal of the oscillator; or (b) to the same terminal. The shield is connected to the center tap of the oscillator output. The quantity K_0 in Eq. (30) is the ratio of the deflections when connections (a) and (b) are made in that order. For the experimental line, K_0 was found to be equal to unity, signifying that each conductor provided a very effective shield from electric fields due to charges on the other line.

This measuring line was rather difficult to operate. The probe couplings for example were not constant for the lengths of the slots. The optimum conditions for line excitation could not be maintained throughout. Due to the several sources of error the experimental data cannot be considered accurate but are included to show representative information.

UNBALANCED RADIATIVE TERMINATIONS

The reflected balanced and unbalanced components of voltages were measured for two types of unbalanced end-coupled antennas, *viz.*, collinear (or extended) antenna, and perpendicular antenna. This is shown in Fig. 4. For various fixed values of y_1 and y_2, "antenna" lengths d_1 and d_2 were varied by inserting threaded sections. The ends of the $\frac{1}{8}$-in. O.D. brass conductors were semispherical in shape. The $\frac{1}{32}$-in. thick 40×42.3-cm aluminum plate minimized the currents on the outside of the slotted line.

In Figs. 5 and 6, Γ_B and Γ_U are plotted for two different structures as a function of "antenna" lengths. These are the *apparent* reflection coefficients of the unbalanced terminations having a reference plane $\lambda/2$ from the aluminum plate. The simultaneous increase in $|\Gamma_U|$ with decrease in $|\Gamma_B|$ implies that additional balanced-mode power is necessary to supply the unbalanced-mode losses. It can be seen that a minimum of $|\Gamma_U|$ occurs when the antenna length is near $\lambda/2$.

Shear Stress in a Turbulent Boundary Layer*

DONALD ROSS AND J. M. ROBERTSON
Ordnance Research Laboratory, The Pennsylvania State College, State College, Pennsylvania
(Received December 9, 1949)

The method initiated by Fediaevsky for evaluating the shear stress distribution in a turbulent boundary layer under the action of an adverse pressure gradient is improved upon. Use is made of a more suitable polynomial and boundary conditions which include the concept that the turbulence which produces the shear at any point has its origin at the surface upstream from its present location. This "history" theory predicts that the transverse shear gradient in the outer portion of the boundary layer remains constant as the boundary layer develops. Comparison with shear distributions measured at the National Bureau of Standards along a simulated airfoil shows good qualitative agreement and appreciable improvement over Fediaevsky's theory.

INTRODUCTION

IN analyzing the flow of a fluid in a diffuser, along the surface of an airfoil, or along the after portion of a streamlined body, the major problem is the evaluation of the effect of the adverse pressure gradient on the flow conditions in the boundary layer. The critical feature of these flows is the occurrence of separation, which results in high losses and other deleterious effects. Despite the common occurrence of this type of incompressible, turbulent boundary-layer flow, the physics of adverse-pressure-gradient effects is still incompletely understood. Undoubtedly, the ultimate solution of this problem will involve a statistical theory of turbulence, but a reasonably useful theory can be developed through the use of the older methods based on the mean motion. The objective of any boundary layer theory is the prediction of the rate of growth of and distribution of the velocity in the boundary layer. As the growth can be found from the von Karman momentum equation if the shapes of the velocity profiles are known, the problem reduces to the prediction of the velocity distributions.

Various researchers have shown that the velocity profiles can be characterized by single form parameters. However, attempts to correlate these parameters with local flow conditions have not been particularly successful. Thus, the methods of analysis developed by Buri[1] and Gruschwitz[2] have only limited application, and they do not appear to be fundamentally sound.[3,4] More recent analyses,[4,5] based on empirical relations for the rate of change of the velocity form parameter, are more successful. In none of these methods, however, is an attempt made to understand the inner mechanics of the flow phenomena.

As the velocity variation in a boundary layer is a result of turbulent shearing motions, a more fundamental understanding of the problem should evolve from a study of shear stress distributions and the origin of the shear. A first attempt at a shear stress theory was made by Fediaevsky[6] in Russia in 1936. He expressed the shear stress as a power series, with the coefficients evaluated in terms of boundary conditions. Comparison of his shear stress profiles with experiment has indicated serious discrepancies; but in spite of this, the method seems so fundamental that the discrepancies may probably be attributed to the boundary assumptions. The shear stress theory presented in this paper, although

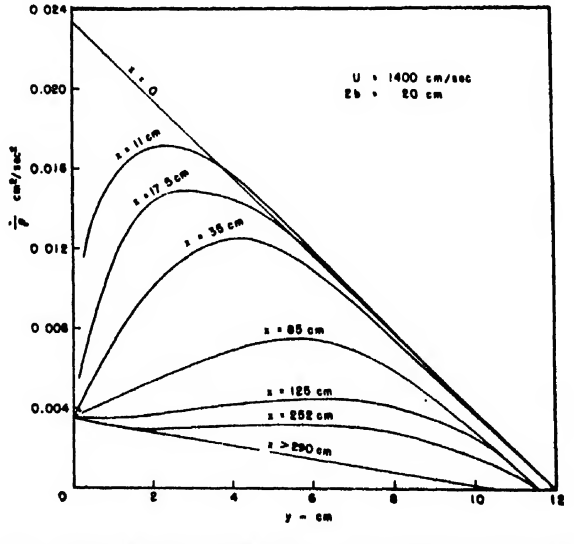


FIG. 1. Shear stress profiles obtained by Jacobs for flow over a smooth surface preceded by a rough surface.

* Work done under Navy Contract NOrd 7958, in connection with the diffuser design for the new water tunnel at The Pennsylvania State College.

[1] A. Buri, "A Method of Calculation for the Turbulent Boundary Layer with Accelerated and Retarded Basic Flow," Thesis No. 652, 1931, Federal Technical College, Zurich Trans. in R.T.P. Translation No. 2073, British Ministry of Aircraft Production.

[2] E. Gruschwitz, Ingenieur-Archiv 2, 321 (1931). E. Gruschwitz, Zeits. f. Flugtechnik und Motorluftschiffahrt 23 (1932), translated in NACA Tech. Memo 699, 1933.

[3] H. Peters, J. Aero. Sci. 3, 7 (1935).

[4] A. E. von Doenhoff and N. Tetervin, "Determination of General Relations for the Behavior of Turbulent Boundary Layers," NACA Report No. 772, 1943 (also NACA Wartime Report L382, ACR No. 3G13, July 1943).

[5] H. C. Garner, "The Development of Turbulent Boundary Layers," ARC R & M 2133, 1944.

[6] K. Fediaevsky, "Turbulent Boundary Layer of an Airfoil," Report 282 Central Aero-Hydrodynamical Institute, Moscow, 1936. Trans. in NACA Tech. Memo 822, 1937, also J. Aero. Sci. 4, 491 (1937).

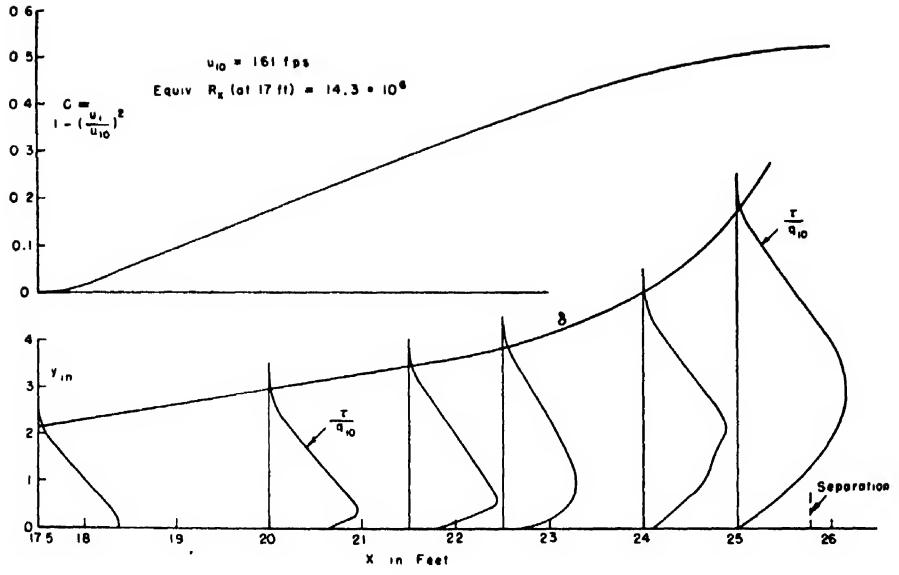


FIG. 2. Shear profiles obtained at the National Bureau of Standards for flow in an adverse pressure gradient. Upper figure shows corresponding pressure regain coefficient. u_1 = velocity just outside boundary layer, u_{10} = initial value of u_1 at start of pressure gradient, $q_{10} = \frac{1}{2}\rho u_{10}^2$.

inspired by Fediaevsky's work, differs from it in several important respects.

FEDIAEVSKY'S THEORY

Fediaevsky[6] used a rational method for determining the form of the velocity distribution for two-dimensional boundary layers from expressions for the shear stress and mixing length distribution. He expressed the transverse form of the shear stress as a power series:

$$\tau/\tau_w = A_0 + A_1(y/\delta) + A_2(y/\delta)^2 + A_3(y/\delta)^3 + \cdots, \quad (1)$$

where τ is the local turbulent shear, τ_w is the wall shear stress, y is the distance from the wall, δ is the disturbance thickness of the boundary layer. He evaluated the coefficients of the first few terms by applying the boundary conditions which must hold at the solid surface and at the outer edge of the boundary layer. At the wall ($y=0$), the function and its first derivative are defined by the following conditions:

(1) $\tau = \tau_w$ by definition of τ_w,
(2) $\partial\tau/\partial y = dp/dx$ from the momentum equation written for the wall streamline.

At the outer edge of the layer ($y=\delta$),

(3) $\tau = 0$ by definition of δ,
(4) $\partial\tau/\partial y = 0$ from the assumption that the derivative with respect to x of the total head is a continuous function.

The derivative of the momentum equation for the wall streamline yields a fifth condition:

(5) $\partial^2\tau/\partial y^2 = 0$ at $y=0$,

By assuming that only the first three conditions are significant, Fediaevsky evaluated the coefficients for a quadratic and obtained

$$\tau/\tau_w = 1 + (\delta/\tau_w)(dp/dx)(y/\delta) - [1 + (\delta/\tau_w)(dp/dx)](y/\delta)^2, \quad (2)$$

while by assuming all five conditions, he obtained the quartic

$$\tau/\tau_w = 1 + (\delta/\tau_w)(dp/dx)(y/\delta) - [4 + 3(\delta/\tau_w)(dp/dx)](y/\delta)^3 + [3 + 2(\delta/\tau_w)(dp/dx)](y/\delta)^4. \quad (3)$$

Although Fediaevsky states that he found fair agreement between his theory and the experimental data of Gruschwitz,[2] Dryden[7] found poor correlation with data obtained at the National Bureau of Standards.

ORIGIN OF SHEAR STRESS

One of the shortcomings of Fediaevsky's theory is its failure to consider the origin of the turbulent motions causing the shear and the ensuing mechanism of transfer. For the case of flow with zero or small pressure gradients, the flow conditions and the wall shear stress do not vary significantly in the direction of flow, and one can correlate the entire boundary layer distribution with local conditions. However, when the flow conditions vary, as in a diffuser or along an airfoil surface, this simple correlation does not hold. The motions which cause the shear forces originate at the wall and are diffused into the fluid stream through the action of turbulence. As this diffusion occurs relative to the moving fluid, the shear at any point in the boundary layer has its origin at a point on the surface upstream from its present location. Thus, the shear in the outer-boundary layer region is governed by the previous history of the flow, while near the solid surface, local conditions control.

[7] H. L. Dryden, "Some recent contributions to the study of transition and turbulent boundary layers," NACA Technical Note 1168 (April, 1947).

TABLE I. History theory assumed shear stress boundary conditions.

Quantity	at $y=0$		at $y=\delta$	
τ	τ_w	(1)	0	(2)
$\partial\tau/\partial y$	dp/dx	(3)	$-(\tau_{w_0}/\delta_0)^*$	(4)
$\partial^2\tau/\partial y^2$	0	(5)	—	

* Differs from Fediaevsky's Theory (subscript 0 refers to initial conditions).

The concept of the effect of space-history on the flow in the outer region of a boundary layer was proposed by Schultz-Grunow[8] and was suggested by the experiments of Jacobs.[9] Jacobs measured the change in the velocity profile along a smooth surface preceded by a rough surface. An analysis of these profiles yielded the shear stress distributions at several stations. As shown in Fig. 1, Jacobs found that the shear stress gradient in the outer portion of the boundary layer retained the value which it had had over the rough surface for a considerable distance along the smooth surface, while the wall shear fell almost immediately to its smooth-surface value. Jacobs also studied the flow from a smooth to a rough surface, and he again found that the upstream shear stress gradient persisted for a considerable distance, the rough-surface shear gradually being diffused across the boundary layer as the flow proceeded downstream. As noted by Schultz-Grunow, these experiments clearly establish the importance of the history of the flow.

Recent studies on the flow along the surface of a large simulated airfoil were made in the wind tunnel at the National Bureau of Standards and were reported by Dryden.[7] In these experiments, the shear stresses were measured more directly with hot wires. The manner in which the NBS shear stress profiles vary in the adverse pressure gradient is shown in Fig. 2. The slope of the

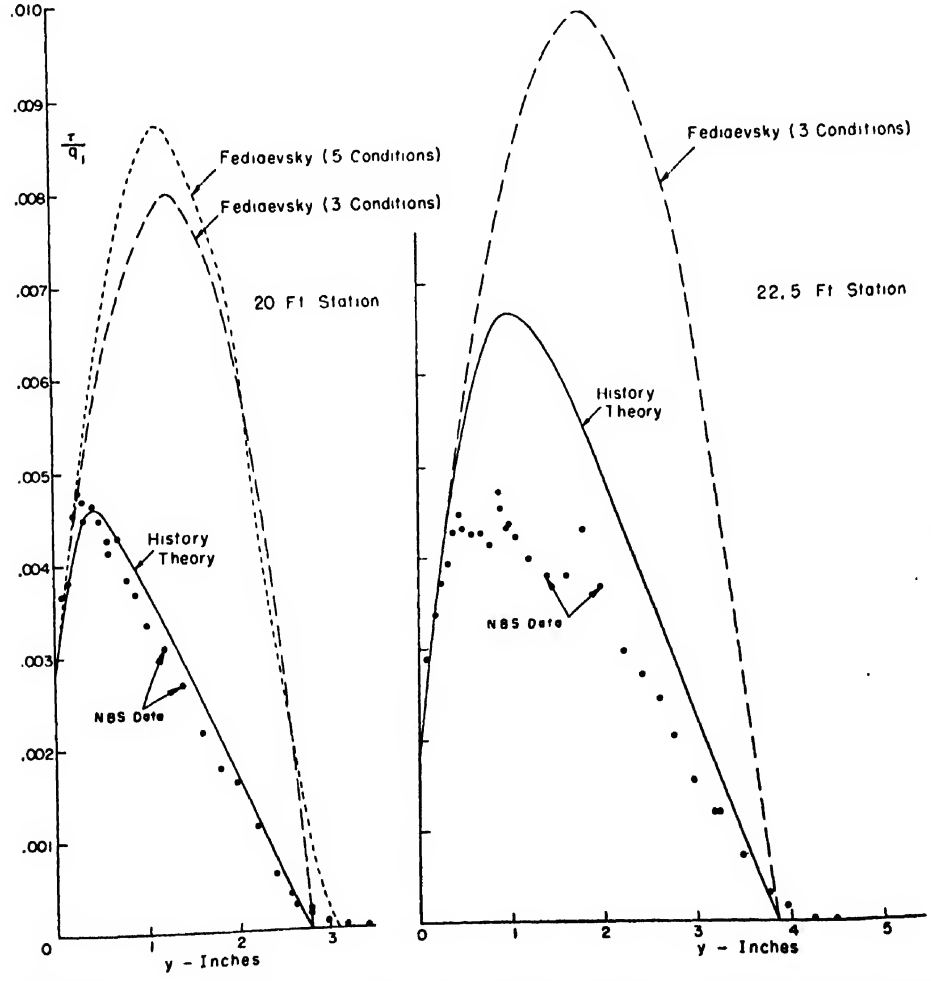


FIG. 3. Comparison of History and Fediaevsky theories with NBS experimental shear measurements for locations near the beginning of the adverse pressure gradient and midway toward separation.

[8] F. Schultz-Grunow, "Über das Nachwirken der Turbulenz bei Örtlich und Zeitlich Verzögerter Grenzschichtströmung," Proc. 5th Int. Cong. Appl. Mech., Cambridge, Massachusetts (1938), pp. 428–435.
[9] W. Jacobs, Zeits. f. angew. Math. u. Mech. 19, (1939). Translated in NACA Tech. Memo 951 (1940).

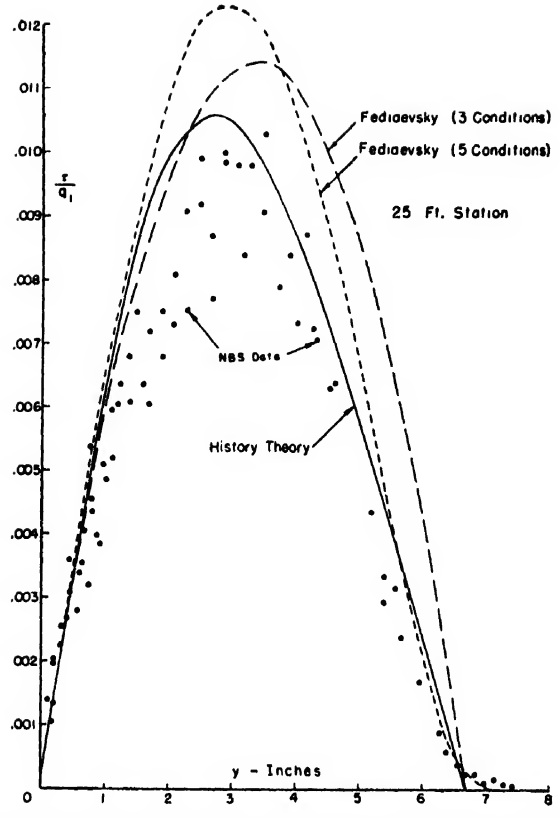


FIG. 4. Comparison of History and Fediaevsky theories with NBS data for station near separation.

shear at the start of the pressure gradient is seen to be carried along in the outer portion of the boundary layer, in qualitative agreement with the history concept.

THE HISTORY THEORY

A theory for the shear stress distribution in a turbulent boundary layer can be derived in a manner similar to that of Fediaevsky but taking into account the space-history effect. Thus, four of the five assumed boundary conditions on the shear stress function presented in Table I agree with those of Fediaevsky. Only the fourth condition, which is for the first derivative of the shear stress function at the outer edge, differs from his assumption. The derivative of the shear stress function is actually zero at the outer limit of the boundary layer, as assumed by Fediaevsky; however, as a polynomial having only a few terms is to be used, the value of the slope over a larger region should be considered. This condition for the slope of the shear stress function at the outer edge of the boundary layer should express the history concept, namely that the slope should have the same value as it had at the beginning of the adverse pressure gradient, where the shear curve is essentially linear.[10] The subscript zeros in condition 4 refer to values at the beginning of the pressure gradient.

As the predominant term in the shear stress function is the linear history term, the shear stress is expressed as a polynomial in terms of the distance from the edge of the boundary layer. To give flexibility to the shape of the curve, while using only a few terms, a function with adjustable exponent has been chosen:

$$\tau = a_0 + a_1[1-(y/\delta)] + a_n[1-(y/\delta)]^n + a_{n+1}[1-(y/\delta)]^{n+1}. \quad (4)$$

Using conditions 1 through 4, the coefficients may be evaluated in terms of the known flow parameters and the exponent n:

$$a_0 = 0,$$
$$a_1 = \tau_{w_0}(\delta/\delta_0),$$
$$a_n = (n+1)\tau_w + \delta(dp/dx) - n\tau_{w_0}(\delta/\delta_0),$$
$$a_{n+1} = -[n\tau_w + \delta(dp/dx) - (n-1)\tau_{w_0}(\delta/\delta_0)].$$

The value of the exponent n may then be fixed from condition 5, yielding

$$n = \frac{2\delta(dp/dx) + \tau_{w_0}(\delta/\delta_0) + \tau_w}{\tau_{w_0}(\delta/\delta_0) - \tau_w}. \quad (5)$$

The resultant expression for the shear stress distribution is

$$\tau = \tau_w(\delta/\delta_0)[1-(y/\delta)] - [\delta(dp/dx) + \tau_{w_0}(\delta/\delta_0)] \times [1-(y/\delta)]^n + [\delta(dp/dx) + \tau_w][1-(y/\delta)]^{n+1}, \quad (6)$$

where n is given by (5). At the start of a region of adverse pressure gradient, the exponent n is very large and the equations yield an essentially linear shear profile. As the boundary layer grows, the exponent decreases and the linear term is relegated to the outer portion of the boundary layer. In this manner, the space-history effect governs the essential characteristics of the shear stress distributions.

EXPERIMENTAL CONFIRMATION

A quantitative check of the history theory can be made by comparison with the shear stress data measured at the National Bureau of Standards[7] and introduced in Fig. 2. Detailed comparisons for stations near the beginning of the adverse pressure gradient, near separation, and midway are shown in Figs. 3 and 4. In each case, the results of the Fediaevsky theory are included for comparison. In these figures, the shear stress is expressed in terms of q_1, the dynamic pressure at the edge of the boundary layer, while in Fig. 2, the initial dynamic pressure q_{10} was used. Near the beginning (20-foot station), the agreement is excellent, and the improvement over Fediaevsky's theory is striking. At the second location (22.5 feet), the history theory is not such a good fit, but is a considerable improvement over the older theory. Although both theories agree reasonably well with experiment near separation (Fig. 4), the

[10] This involves a discontinuity in the shear stress and velocity distributions at the outer edge of the boundary layer. As in the case of the turbulent boundary layer along a flat plate, this discontinuity does not introduce any gross discrepancies.

history theory offers only a small improvement. From this limited comparison, it is apparent that the history theory gives reasonable agreement with experiment, especially in the beginning of the adverse pressure gradient where the older theory was most unsatisfactory.

In obtaining the theoretical shear stress distributions shown in Figs. 3 and 4, experimental values of the boundary layer thickness and the wall shear stress had to be used. Near the beginning of the pressure gradient, the wall shear has essentially the flat-plate value, whereas at separation the value is zero. For completeness, a theory is needed which will yield the local wall shear in terms of its initial value and the other flow parameters.

SUMMARY

A satisfactory theory for the shear stress distribution in a turbulent boundary layer has been developed from consideration of the space-history of the turbulence. This theory evolved from Jacobs' observations that the shear stress gradient in the outer portion of the boundary layer is constant, even when the external flow conditions vary in the direction of flow. Following Fediaevsky's method, boundary values are used to determine the coefficients of an arbitrary polynomial. The theory yields a linear shear profile at the beginning of the adverse pressure gradient; and as the boundary layer progresses, this linear characteristic is relegated to the outer part and the shear in the inner region is determined by local conditions. Comparison with shear measurements made at the National Bureau of Standards shows good agreement and appreciable improvement over Fediaevsky's theory. A more complete theory of the shear variation in a boundary layer under an adverse pressure gradient should consider turbulence decay and should also evaluate the wall shear stress.

Propagation of the TM_{01} Mode in a Metal Tube Containing an Imperfect Dielectric*

DAVID L. HETRICK

Department of Physics, Rensselaer Polytechnic Institute, Troy, New York

(Received September 20, 1949)

The TM_{01} mode in a circular dielectric-filled wave guide, where dielectric losses and metal wall losses are considered simultaneously, is analyzed from the viewpoint of the boundary value problem. General expressions for the attenuation and phase constant are obtained, and the percentage errors introduced by (1) neglect of wall losses and (2) assumption that the total attenuation is the sum of dielectric attenuation and approximate wall loss attenuation are computed for a particular case.

INTRODUCTION

THE problem of the propagation of electromagnetic waves in a uniform metal tube has usually been solved for the case where the metal of the tube itself is approximated by a perfect conductor. Even when losses in the enclosed dielectric are considered by retaining the imaginary part of the complex dielectric constant, the analysis does not become difficult.

When losses arising from the penetration of the fields into the metal sheath are to be taken into account, two approaches are available. One may assume that the field configurations are not appreciably altered from those obtained with perfect walls, and, knowing the intrinsic impedance of the bounding medium, compute the power loss and attenuation. The total attenuation is then taken to be the sum of attenuations arising from dielectric losses and wall losses separately.

The more general approach is the boundary value problem wherein the continuity of tangential fields leads to expressions for the attenuation and phase constant. Again, certain approximations are necessary, but these have negligible effect on the accuracy of the results. Insofar as the author is aware, the solution as a boundary value problem has not been carried out previously where dielectric losses and wall losses are considered simultaneously. The object of this investigation is to obtain general expressions for the attenuation and phase constant and to show for a particular case how the results differ from those obtained when (1) the wall losses are neglected and (2) the total attenuation is taken to be the sum of dielectric attenuation and approximate wall loss attenuation. For simplicity, only the circularly symmetric TM mode is considered.

The method followed was first employed by Carson, Mead, and Schelkunoff,[1] and by Barrow.[2] This phase of the problem was investigated in more detail by Linder,[3] though only the wall losses for an air-filled guide were considered in these papers. Recently[4–8] several papers

* Based on a thesis submitted to the Department of Physics of Rensselaer Polytechnic Institute in partial fulfillment of the requirements for the M.S. degree.

[1] Carson, Mead, and Schelkunoff, Bell Sys. Tech. J. **15**, 310–333 (1936).
[2] W. L. Barrow, Proc. I.R.E. **24**, 1298–1329 (1936).
[3] E. G. Linder, Proc. I.R.E. **30**, 554–556 (1942).
[4] W. P. Conner and C. P. Smyth, J. Am. Chem. Soc. **65**, 382–389 (1943).
[5] J. G. Powles, Trans. Faraday Soc. **44**, 537–545 (1948).
[6] W. H. Surber, Jr., J. App. Phys. **19**, 514–523 (1948).
[7] Heston, Hennelly, and Smyth, J. Am. Chem. Soc. **70**, 4093–4096 (1948).
[8] H. L. Laquer and C. P. Smyth, J. Am. Chem. Soc. **70**, 4097–4101 (1948).

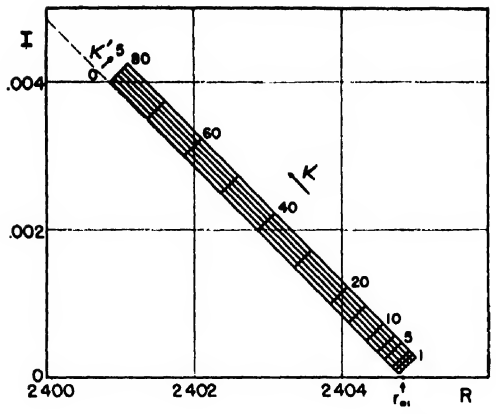


FIG. 1. Curves of constant κ and κ' in the complex ρ_1 plane; $\lambda_0/a=2\pi$, and a free-space wave-length of 1 cm is assumed.

have been published discussing the use of wave guide methods in measuring complex dielectric constants. It is important to know how great an error is introduced into the results when the wall losses are neglected, particularly if these results are to be used in deducing the structures of complex organic molecules.

THE BOUNDARY EQUATION

The fields whose continuity must be preserved at the walls of the wave guide may be expressed as[9] (MKS units)

$$E_{1z}=Ak_1^2J_0(k_1r)\exp(j\omega t-\gamma z), \tag{1}$$

$$E_{2z}=Bk_2^2H_0^{(1)}(k_2r)\exp(j\omega t-\gamma z), \tag{2}$$

$$H_{1\phi}=j\omega\epsilon_{c1}Ak_1J_1(k_1r)\exp(j\omega t-\gamma z), \tag{3}$$

$$H_{2\phi}=j\omega\epsilon_{c2}Bk_2H_1^{(1)}(k_2r)\exp(j\omega t-\gamma z), \tag{4}$$

where

$$k^2=\gamma^2-j\omega\mu(\sigma+j\omega\epsilon), \tag{5}$$

and the propagation constant γ is

$$\gamma=\alpha+j\beta=\alpha+j(2\pi/\lambda).$$

The subscripts 1 and 2 refer to the enclosed dielectric and the metal sheath respectively. These are taken to be regions containing no free charges and having conductivity σ, permittivity ϵ, and permeability μ; α is the attenuation constant, β the phase constant, λ the guide wave-length, ω the angular frequency $2\pi\nu$, A and B arbitrary constants, $J_n(kr)$ and $H_n^{(1)}(kr)$ the Bessel function and Hankel function of the first kind respectively. The complex permittivity ϵ_c is defined as

$$\epsilon_c=\epsilon-j\sigma/\omega. \tag{6}$$

Cylindrical co-ordinates r, ϕ, and z are employed, where z is the direction of the wave guide axis.

Note that $\gamma_1=\gamma_2$ in order to satisfy the boundary conditions for all z and t. The argument kr must be taken as having a positive imaginary part; it can be shown that this leads to no loss of generality.

With the use of Eq. (5), assuming $\mu_1=\mu_0$,

$$k_1=[\gamma^2+(2\pi/\lambda_0)^2(\kappa-j\kappa')]^{\frac{1}{2}} \tag{7}$$

where $\kappa=\epsilon/\epsilon_0$, $\kappa'=\sigma_1/\omega\epsilon_0$, and λ_0 is the free-space wave-length. In the metal, $\omega\mu_2\sigma_2\gg\gamma^2$ and $\omega\mu_2\sigma_2\gg\omega^2\mu_2\epsilon_2$. Assuming $\mu_2=\mu_0$,

$$k_2=2\pi/\lambda_0[-j\sigma_2/\omega\epsilon_0]^{\frac{1}{2}}. \tag{8}$$

Setting $\rho_1=k_1a$ and $\rho_2=k_2a$, where a is the inner radius of the metal sheath, one obtains from the continuity of tangential fields at the boundary

$$\frac{\epsilon_{c1}J_1(\rho_1)}{\rho_1J_0(\rho_1)}=\frac{\epsilon_{c2}H_1^{(1)}(\rho_2)}{\rho_2H_0^{(1)}(\rho_2)}. \tag{9}$$

This condition is sufficient to determine the constants of propagation. In general, a superposition of *TM* and *TE* waves is necessary to satisfy the boundary conditions, but in the circularly symmetric case no new mode is introduced.[9]

The values of ρ_2 justify the use of the asymptotic expansion[10]

$$H_1^{(1)}(\rho_2)/H_0^{(1)}(\rho_2)=e^{-j\pi/2}. \tag{10}$$

Replacing ϵ_{c1}/ϵ_0 by κ_c, and making use of Eqs. (6), (8),

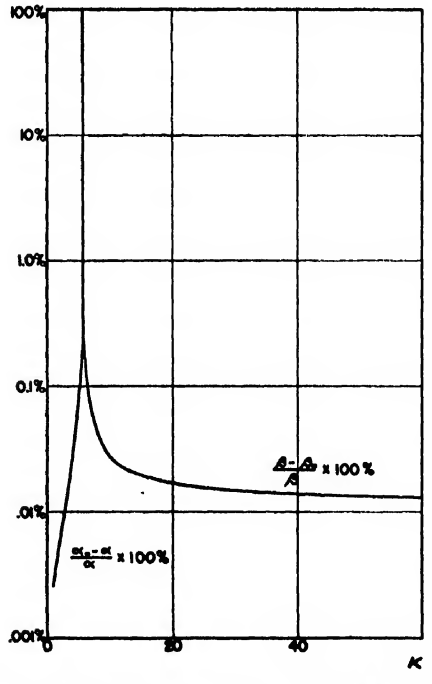


FIG. 2. Percentage errors in attenuation and phase constant introduced when metal wall losses are neglected; $\lambda_0/a=2\pi$ and $\kappa'=0$.

[9] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), p. 521 *et seq.*

[10] R. I. Sarbacher and W. A. Edson, *Hyper and Ultrahigh Frequency Engineering* (John Wiley & Sons, Inc., New York, 1943), p. 231 *et seq.*

and (10), Eq. (9) becomes

$$\frac{\kappa_c J_1(\rho_1)}{\rho_1 J_0(\rho_1)} = \frac{\lambda_0}{2\pi a}\left[\frac{\sigma_2}{2\omega\epsilon_0}\right]^{\frac{1}{2}}(1+j). \tag{11}$$

Since $\kappa_c = \kappa - j\kappa'$, divide by κ_c to obtain

$$J_1(\rho_1)/\rho_1 J_0(\rho_1) = c + jd \tag{12}$$

where

$$c = \frac{\lambda_0}{2\pi a}\left[\frac{\sigma_2}{2\omega\epsilon_0}\right]^{\frac{1}{2}}\frac{\kappa-\kappa'}{\kappa^2+\kappa'^2} \tag{13}$$

and

$$d = \frac{\lambda_0}{2\pi a}\left[\frac{\sigma_2}{2\omega\epsilon_0}\right]^{\frac{1}{2}}\frac{\kappa+\kappa'}{\kappa^2+\kappa'^2}. \tag{14}$$

In order to solve Eq. (12) for ρ_1, replace the Bessel functions by the first two terms of their complex Taylor series expansions

$$J_1(\rho_1) = J_1(r_{01}) + (\rho_1 - r_{01})J_1'(r_{01}) + \cdots,$$
$$J_0(\rho_1) = J_0(r_{01}) + (\rho_1 - r_{01})J_0'(r_{01}) + \cdots,$$

where $r_{01} = 2.404,825,558$, the root[11] of rank one of $J_0(k_1 a) = 0$, and the limiting value of ρ_1 as σ_2 becomes infinite.[10] This yields expressions for the TM_{01} mode.

One obtains a quadratic

$$[c+jd]\rho_1^2 - [(r_{01}c + 1/r_{01}) + jr_{01}d]\rho_1 + 2 = 0 \tag{15}$$

having the solution

$$\rho_1 = R + jI, \tag{16}$$

where

$$R = \frac{r_{01}}{2} + \frac{c}{2r_{01}(c^2+d^2)} + \left[\frac{L}{2} + \frac{1}{2}(L^2+M^2)^{\frac{1}{2}}\right]^{\frac{1}{2}}, \tag{17}$$

$$I = -\frac{d}{2r_{01}(c^2+d^2)} + \left[-\frac{L}{2} + \frac{1}{2}(L^2+M^2)^{\frac{1}{2}}\right]^{\frac{1}{2}}, \tag{18}$$

$$L = \frac{r_{01}^2}{4} + \frac{c^2 - d^2 - 6cr_{01}^2(c^2+d^2)}{4r_{01}^2(c^2+d^2)^2}, \tag{19}$$

$$M = \frac{3dr_{01}^2(c^2+d^2) - dc}{2r_{01}^2(c^2+d^2)^2}, \tag{20}$$

and c and d are defined by Eqs. (13) and (14).

The ambiguity in sign arising in the solution of the quadratic is removed by observing the effect of allowing σ_2 to increase without limit. The sign has been chosen so that $\rho_1 \to r_{01}$; the alternative leads to the impossibility $\rho_1 \to 0$.

Figure 1 is a typical set of curves of constant κ and κ' in the complex ρ_1 plane showing the departure from the case of perfect walls. The curves are for a copper tube, assuming a free-space wave-length of 1 cm and a ratio $\lambda_0/a = 2\pi$. Note that the curve for $\kappa' = 0$, when extended, passes through the point $R = r_{01}$, $I = 0$. For $I < 0.002$, the curves of constant κ' approach straight lines.

ATTENUATION AND PHASE CONSTANT

From Eq. (7),

$$\rho_1^2 = (R+jI)^2 = a^2[(\alpha+j\beta)^2 + (2\pi/\lambda_0)^2(\kappa - j\kappa')].$$

Equating real and imaginary parts, a quadratic in $(a\alpha)^2$ is obtained. Choosing only positive real values for α, one obtains

$$a\alpha = 2^{-\frac{1}{2}}[q_1 + (q_1^2+q_2^2)^{\frac{1}{2}}]^{\frac{1}{2}}, \tag{21}$$

where

$$q_1 = R^2 - I^2 - (2\pi a/\lambda_0)^2\kappa,$$
$$q_2 = 2RI + (2\pi a/\lambda_0)^2\kappa'.$$

The phase constant is given by

$$a\beta = q_2/2a\alpha. \tag{22}$$

Equations (21) and (22) will be convenient for numerical computations only in the "region of attenuation" where $(\lambda_0/a)\kappa^{-\frac{1}{2}} > 2\pi/r_{01}$. In the "region of propagation" a quadratic in $(a\beta)^2$ yields

$$a\beta = 2^{-\frac{1}{2}}[-q_1 + (q_1^2+q_2^2)^{\frac{1}{2}}]^{\frac{1}{2}}. \tag{23}$$

Once the phase constant is known from Eq. (23), Eq. (22) may be employed to obtain the attenuation.

In the limiting case of perfectly conducting walls where $R \to r_{01}$ and $I \to 0$, these general expressions yield the attenuation and phase constant for the case of dielectric loss only. If, in addition, $\kappa' \to 0$, one obtains the expressions for the completely lossless case.[10]

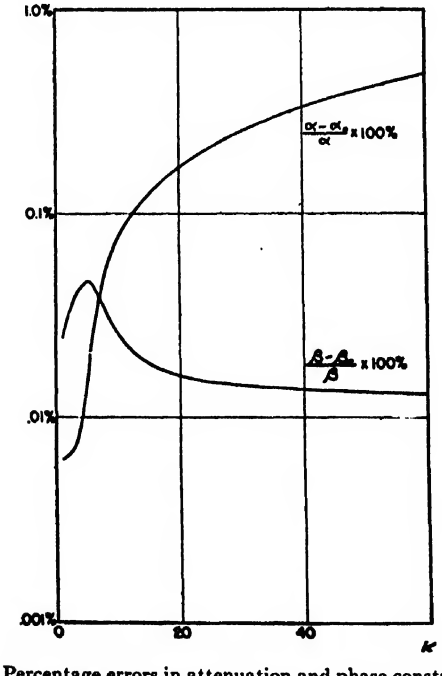


FIG. 3. Percentage errors in attenuation and phase constant introduced when metal wall losses are neglected; $\lambda_0/a = 2\pi$ and $\kappa' = 3$.

[11] Gray, Mathews, and MacRobert, *A Treatise on Bessel Functions* (MacMillan & Company, London, 1922), p. 244.

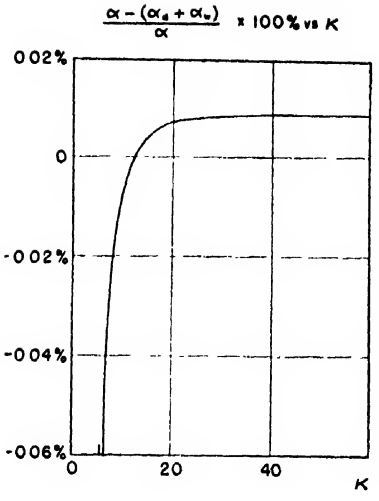


FIG. 4. Percentage error in attenuation introduced by assuming $\alpha=\alpha_d+\alpha_w$; $\lambda_0/a=2\pi$ and $\kappa'=3$.

NUMERICAL RESULTS

The percentage errors in attenuation and phase constant introduced by neglecting wall losses are shown in Figs. 2 and 3, where $\kappa'=0$ and 3 respectively. The subscript 0 on α and β denotes the completely lossless case, and the subscript d denotes values obtained from Eqs. (21) through (23) for dielectric loss only. Values obtained from these equations when wall losses are included are indicated without subscripts. The critical point, or transition from attenuation to propagation for the lossless guide, is found at $\kappa=5.78$ for the case chosen.

The errors arising from the assumption that the total attenuation is $\alpha_d+\alpha_w$, where α_w represents the approximate wall loss attenuation as computed from power flow considerations,[10] are shown in Fig. 4.

Two remarkable effects will be noted. In Fig. 2, where no dielectric loss is present, the attenuation decreases when wall losses are added; $\alpha_0>\alpha$. This seems contrary to expectation at first glance. A qualitative explanation may be offered as follows: α_0 is a purely reactive attenuation. When wall losses are added, the penetration of the fields into the metal causes an effective change in the field configuration which reduces the reactive attenuation. This is evidently more than enough to compensate for the resistive attenuation caused by the metal walls.

The second interesting effect is that the error curve in Fig. 4 changes sign; there is one value of κ for which $\alpha=\alpha_d+\alpha_w$ exactly.

While these curves apply to a specific set of parameters, one may conclude in general that the assumption $\alpha=\alpha_d+\alpha_w$ is an excellent approximation anywhere in the "region of propagation" except in the immediate neighborhood of the critical point; the relative error in attenuation remains of the order of magnitude of 10^{-4} for most of the range of operation. On the other hand, the wall losses cannot be entirely neglected in this range of operation where relative errors less than 10^{-3} are desired. The order of magnitude of the errors involved will not vary greatly as the cross-sectional configuration of the wave guide is altered.

DIELECTRIC CONSTANT AND POLAR MOLECULES

It can be shown[12] that if the real and imaginary parts of the dielectric constant of a polar liquid are known, much information may be gained about the molecular structure of the liquid. For example, the polarizability, dipole moment, and relaxation time of the dipoles may be calculated if κ and κ' are known with sufficient accuracy.

The κ' referred to here arises from polarization losses only. If ionic conduction is present, its effect must be subtracted.

The values of κ and κ' may be obtained from wave guide measurements of attenuation and phase constant in the following way: Suppose that curves of $a\alpha$ and $a\beta$ *vs.* κ for various κ' are available.** If horizontal lines are drawn on each set of curves corresponding to measured values of $a\alpha$ and $a\beta$ respectively, data for curves of constant $a\alpha$ and $a\beta$ in the complex κ_c plane are obtained. The intersection determines κ and κ'.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Professor Philip Rosen of the Department of Physics of Rensselaer Polytechnic Institute for his suggestion of this investigation and for his assistance in its completion. Particular thanks are due to Professor Alfred W. Jones of the Department of Mathematics and to Professor Robert M. Whitmer of the Department of Physics for their invaluable aid.

[12] Glasstone, Laidler, and Eyring, *The Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941), pp. 544–546.

** It is not necessary that the only term in κ', as used in preceding sections, be the term $\sigma_1/\omega\epsilon_0$. The introduction of a polarization loss in the dielectric does not alter the wave guide equations.

Ultra Speed Tensile of Rubber and Synthetic Elastomers*

D. S. VILLARS†
General Laboratories, United States Rubber Company, Passaic, New Jersey
(Received October 6, 1949)

A high speed stress-strain machine has been developed which is capable of recording the stress-strain curve of elastomers at elongation rates up to 270 percent/msec. Data are reported on two series of gum and tread stocks of hevea and of the synthetic elastomers, GR-S, Hycar OR, butyl, Perbunan, and Neoprene GN. The second (elastomer) series was also run at 150°C. In general, stress-strain curves fall into two classes. Stocks of elastomers which are known to crystallize on stretching tend to show tensiles which decrease with increasing speed up to about 10 percent/msec., pass through a minimum and rise more or less drastically to values 100 percent (or more) greater than the Scott tensile. Elastomers which do not crystallize on stretching tend to show a steady rise in tensile with increasing speed. Elongation at break curves show a maximum with crystallizing stocks and no maximum with non-crystallizing stocks. The shape of the modulus *versus* speed curves is accounted for on the hypothesis of different types of slipping bonds with different characteristic relaxation times. The shift of curves for tread stocks with temperature allows the estimation of a heat of activation of slippage. This comes out to be of the order of 3 kg cal.

INTRODUCTION

A MACHINE has been developed which is capable of recording the stress-strain curve of rubber at rates of elongation up to 9000 times that used in the Scott tensile test. This is over 100 times faster than any rate previously attained in any rubber tests. Availability of such high speed, which shall hereinafter be designated as "ultra speed," enables one now, for the first time, to study the relationship to service behavior of fundamental physical properties of commercial rubber stocks. The present paper presents a brief description of the machine and surveys the speed dependence of tensile strength, elongation at break, modulus and breaking energy of rubber and synthetic elastomers.

Other investigators have attempted to measure short time stress strain and tensile of rubber. Albertoni[1] described a pendulum impact tester which measured energy expended to stretch individual samples to definite predetermined elongations and to break. Roth and Holt[2] built a clever guillotine type of machine which stretched samples by means of a vertically falling weight. Dart, Anthony, and Wack[3] devised a fast stress-strain machine working on "the principle of a mechanical balance."

Objections to the foregoing types of machine are: (1) Time of break is not short enough. (2) Too much calculation is required to obtain the stress-strain curve (excessive calculation precludes large volume of investigation). (3) Rate of elongation in Roth-Holt machine is not uniform. (4) A different sample is tested for each point on the stress-strain curve when using the Albertoni machine.

Our present new machine overcomes each of these objections. It gives times of break down to $\frac{1}{2}$ msec. with constant rates of elongation almost 100 times that of the Albertoni machine, 200 times the mean rate of the Roth and Holt apparatus, 1000 times that obtained with the highest frequency dynamic hysteresis apparatus (Naunton and Waring,[4] *ca.* 1000 cycles/sec.) and 9000 times that of the Scott tensile machine. It gives direct photographs of the complete stress-strain curve of a particular sample, without any reduction by integration or differentiation.

Description of Ultra Speed Machine

The machine (cf. Fig. 1) consists of a rotating wheel, on the circumference of which a pin strikes the middle enlarged part of a double dumbbell test piece hanging in the form of a loop. This dumbbell was designed especially for this machine. The two narrow parts are 0.1 in. wide and 17/64 in. long. The total length of test piece is $5\frac{7}{32}$ in. The stress produced is communicated mechanically to a piezoelectric crystal which generates an electrical potential directly proportional to the stress. This potential is amplified and recorded on an oscillograph screen. The oscillograph trace, which is the

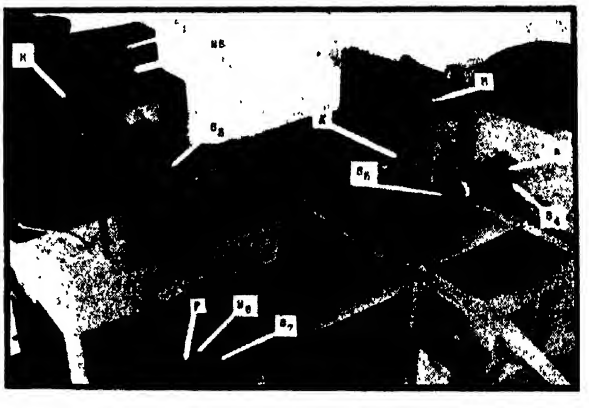


FIG. 1. Ultra speed machine.

* Presented before the Division of Rubber Chemistry at the Atlantic City meeting of the American Chemical Society, September, 1949.

† Present address: U. S. Naval Ordnance Test Station, China Lake, California.

[1] G. J. Albertoni, Ind. Eng. Chem. Anal. **9**, 30–4 (1937); R.C.T. **10**, 317 (1937).

[2] F. L. Roth and W. L. Holt, J. Research Nat. Bur. Stand. **23**, 603–16 (1939); R.C.T. **13**, 348–60 (1940).

[3] Dart, Anthony, and Wack, Phys. Rev. **69**, 52(A) (1946).

[4] Naunton and Waring, Proc. Rubber Tech. Conf. London, 805–20 (1938).

TABLE I.

	ADJ	ADK	ADL	ADM
Pale crepe	100	100		
GR-S			102	102
Blended black		47		46
Stearic acid	2	5	1.5	1.5
XX zinc oxide	5	8	5	5
Antioxidant[a]	1	1		
Captax[b]	1	1	1 5	1.5
Sulfur	2 5	2 5	1 5	1.5
Cure	35 min. @ 30 lb.	30 min. @ 40 lb.	30 min. @ 50 lb.	45 min. @ 50 lb.
	ADN	ADO	ADP	ADQ
Hycar OR	100	100		
GR-I			101.5	101.5
Blended black		44		46
Stearic acid	1	1	3	3
Zinc oxide	5	5	5	10
Age rite powder	1	1		
Altax[c]	1	1		
Tuex[d]			1	1
Sulfur	1	1	0.5	1.5
Cure	30 min. @ 50 lb	30 min. @ 50 lb.	30 min @ 50 lb.	45 min. @ 55 lb.
	ADR	ADS	ADT	ADU
Perbunan	100	100		
Neoprene GN			100	100
Blended black		44		35
Stearic acid	1 5		0.5	0 5
Lauric acid		1.5		
Zinc oxide	5	5	1	1
Extra lt. cal. mag.			4	4
552 acc.[e]			0.5	1.5
Neozone A[f]			2	2
Altax[c]	1	1.25		
Sulfur	1.2	1.25		
Cure	15 min. @ 50 lb.	45 min. @ 50 lb.	15 min. @ 50 lb.	30 min. @ 45 lb.

[a] Reaction product of acetone and diarylamine.
[b] 2-mercapto benzothiazole.
[c] Bis-2-thiobenzothiazole.
[d] Tetramethyl thiuram disulfide
[e] Piperidine salt of *N,N*-pentamethylene dithiocarbamic acid.
[f] Phenyl-β-naphthylamine.

stress-strain curve, is photographed on 35-mm film. Calibration is carried out by suddenly applying a steady force upon the crystal by means of an electromagnet. The strength of the latter is in turn determined by weighing the force necessary to break the armature away from the magnet. By means of a variable speed transmission, and speed reducing pulleys, all wheel speeds between 60 and 1700 r.p.m. can be obtained. The corresponding elongation rates in rubber samples vary from 9.3 to 270 percent/msec. The output of the machine is 13 to 16 stress-strain curves per day.

It is planned to submit a detailed technical description of the machine to *The Review of Scientific Instruments*.

Stocks

Two series of stocks are considered in this paper. In the first, called the "Replication series," standard tread stocks, containing 50 parts carbon black (per hundred of elastomer), and standard gum stocks, omitting all carbon black, were compounded and cured. The second series is of six different elastomers, without and with 21 volumes channel black per 100 volumes matrix (rubber plus soluble components). Table I gives formulas of the second series.

RESULTS

Shape of Stress-Strain Curve

A typical ultra speed stress-strain curve taken at 57 percent/msec. is reproduced in the upper half of Fig. 2. It is approximately linear. Figure 3 plots the average of three stress-strain curves of a rubber gum stock taken at 36 percent/msec. At highest ultra speeds, the stress-strain curves assume a wave form (lower half of Fig. 2), sometimes multiple in character. The reason for this multiplicity is not known. It may have something to do with slippage of the samples in the holders. It may be associated with resonance chattering oscillations in the crystal holder or crystal itself. That it is not due to each wave representing individual non-simultaneous breaks of the two sample sides was shown by the fact that samples with one side cut beforehand also showed double loops.

Velocity of Sound

At the greatest ultra speeds the possibility has been taken into consideration of perturbing effects which might be introduced by the wave character of stress response whenever the rate of elongation approaches too closely the velocity of sound in the sample. For zero hysteresis the stress recorded at the held end of a uniform sample elongated at a constant rate should increase with time by finite perpendicular steps.[5] Al-

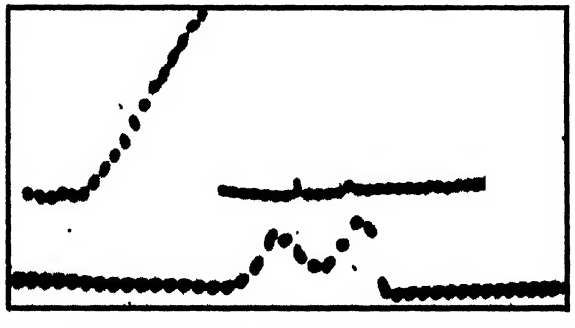


FIG. 2. Ultra speed stress strain curves. Upper curve: rubber tread; elongation rate: 57 percent/msec.; timing frequency: 2000 c.p.s.; tensile: 3530 lb./in^2.; elongation at break: 325 percent; time of break: 5.7 msec. Lower curve: special rubber tread; elongation rate: 238 percent/msec.; timing frequency: 5000 c.p.s.; tensile: 6600 lb./in^2.; elongation at break: 143 percent; time of break: 0.6 msec.

[5] Fourier analysis of the equation

$$(\partial^2 y/\partial \tau^2) = (\partial^2 y/\partial \xi^2)$$

for the boundary conditions,

$$\begin{aligned} y &= 0 \quad \text{at } \tau = 0, \\ (\partial y/\partial \tau) &= 0 \quad \text{at } \tau = 0, \\ y &= 0 \quad \text{at } \xi = 0, \\ y &= \alpha\tau \quad \text{at } \xi = 1, \ \tau > 0 \end{aligned}$$

though we know that we are dangerously near the velocity of sound at highest speeds, no elastomer curves have yet been observed which have the step shape predicted by wave theory. This fact indicates that none of our conclusions are subject to modification because of wave phenomena. However, it is still possible that large specific acoustic absorption effects at the frequencies concerned might reduce the steepness of the "steps." This could result in the steps being overlooked.

Preworking

Possibility of an effect of preworking was checked by running a series of rubber tread stocks, half of the samples of which were preworked by (hand) stretching ten times out to the "hardening" point and back. The order of testing these and unworked controls was random. Results (Table II) indicate no great effect on either tensile or elongation.

Comparison of High and Low Speed Stress-Strain Curves

Figure 3 compares the average of three ultra speed curves (elongation rate 36 percent/msec.) of rubber gum stock with a slow speed stress (single pull—no pre-

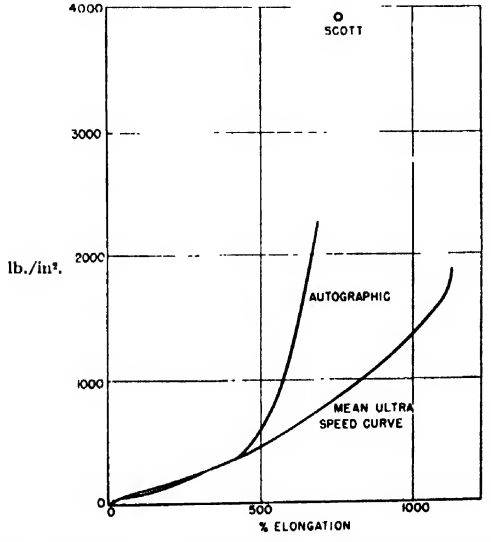


FIG. 3. Comparison of ultra speed with autographic stress-strain curve—rubber gum. Times of break—ultra speed: 28 msec.; Scott: 30 sec.; autographic: 88 sec.

gives:

$$y = \alpha\xi\tau + \frac{2\alpha}{\pi^2}\sum_{m=1}^{\infty}\frac{(-1)^m}{m^2}\sin m\pi\xi\,\sin m\pi\tau.$$

In these equations the dimensionless quantities are defined as:

$$y = u/L;\quad \xi = x/L;\quad \tau = t/(L/v);\quad \alpha = a/v,$$

where u is the displacement of a point in the sample of total length L, originally at a distance x from the fixed end, t is time, a is the uniform velocity of displacement of the other end of the sample and v is the velocity of sound in the rubber (taken as constant). The local elongation $\partial y/\partial\xi$ at $\xi = 0$ (and hence stress) may be shown to represent a series of perpendicular steps, each of height 2 dimensionless units and of length 2 units of τ.

TABLE II. Effect of preworking rubber tread. Mean time of break, 8 msec.

Tensiles		Elongations	
Unworked	Preworked	Unworked	Preworked
2910	3180	442	540
2790	2610	442	388
	2010		334
Mean 2850	2600	442	421

TABLE III. Testing error of ultra speed tensile—quartz crystal. Tread stocks—AKE-I.

Elongation rate percent/msec.	Standard deviation, lb./in.² mean of five
11	80
24	120
55	170
120	305
260	370

working) strain curve of a sample from the same slab of rubber, as recorded by an autographic stress strain machine designed by H. M. Smallwood, H. T. Battin, and W. Moorman. This machine uses ring-shaped test pieces, 1.373 in. o.d. and 1.173 in. i.d. Such test pieces never give the full value of the tensile strength, probably "because of uneven distribution of stress over the cross section caused by the fact that the inside of the ring is stretched more than the outside during elongation and because of the effect of the clamp pulleys on the portions of the ring passing over them."[6] However, the autographic curve, Fig. 3, is seen to head for the Scott tensile value, also determined on pieces from the same slab. It breaks away from the ultra speed curve at about 400 percent elongation.

Reproducibility of Measurement

To obtain an accurate estimate of the variability of measurement, triplicate mixes of Hevea and GR-S gum and tread stocks were mixed in random order and cured. Each test slab was $6\frac{1}{2}\times 6\frac{1}{2}\times 0.1$ in. and can supply five ultra speed stress-strain, four Scott, and one autographic stress-strain test pieces. In this test, five speeds were investigated and each mix-speed combination was run in random order until each combination was represented. Such a "block" was then repeated in a new random order until five blocks were completed.

In carrying out such a program plan, stress-strain curves are occasionally missed. This is sometimes caused by faulty synchronization, sometimes by failure of the sweep trigger mechanism, not to speak of human failures to open the camera shutter or turn the film on to the next frame. Such missing values were filled in by the method of Yates.[7]

[6] Arthur W. Carpenter in Davis and Blake, *The Chemistry and Technology of Rubber* (Reinhold Publishing Corporation, New York, 1937), p. 791.

[7] F. Yates, "Analysis of replicated experiments where . . . results are incomplete," Empire J. Exper. Agric. 1, 129–42 (1933).

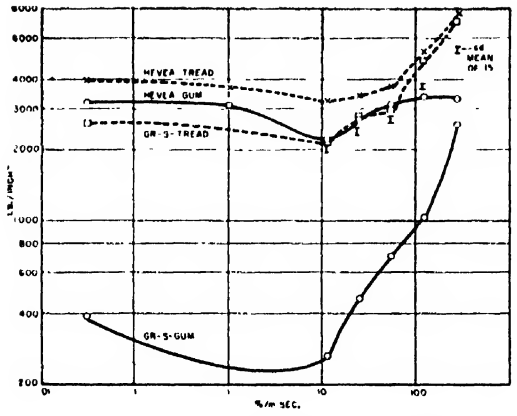


FIG. 4(a). Tensile—Replication series AEG-J.

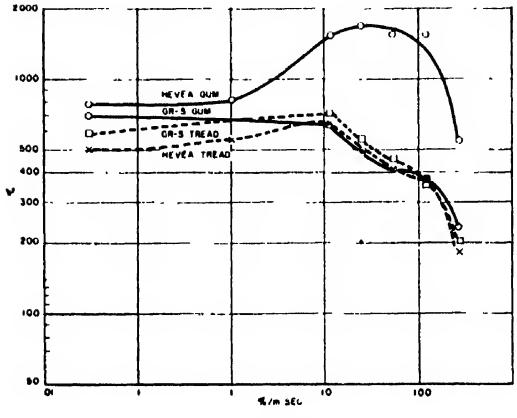


FIG. 4(b). Elongation at break—Replication series AEG-J.

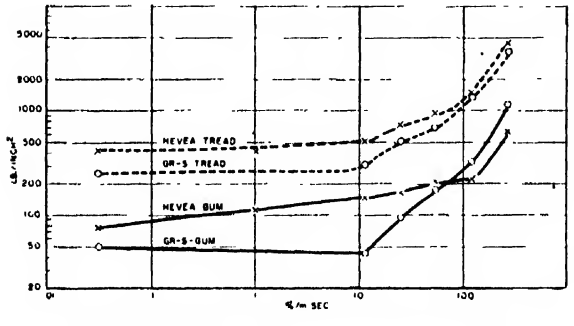


FIG. 4(c) Modulus—Replication series AEG-J.

The standard deviations of testing error were determined separately for each individual rate of elongation. See Table III. It may be seen from the I-shaped marks on Fig. 4(a) that the standard deviation of tensile strength tends to vary in direct proportion with the value of the tensile (coefficient of variation is more constant than the standard deviation).

No mixing error (significantly greater than testing error) was found except at speed 120 percent/msec. This means that testing error is the predominant source of variability.

TENSILE STRENGTH

Dependence on Speed

Figure 4(a) plots the results of ultra speed tensile measurements on the first (replication) series of Hevea gum and tread and GR-S gum and tread. The abscissae are logarithms of elongation rates. At the left of the graph is the Scott tensile value, plotted at 0.03 percent/msec. Each of the ultra speed values (in this series only) is the mean of five samples each of three independent mixes of the same stock and the standard deviations, listed in Table III as for means of 5, are indicated by the limits of the I-shaped figures on the graphs (after converting to means of 15). From these limits one may judge visually the significance of the trends exhibited.

Roth and Holt[2] values for Hevea are inserted at 1 percent/msec. Their tread compound contained 40 parts channel black, while ours contained 50 parts. That particular point falls reasonably on our curve, as if variation in carbon black loading in that range makes little difference at that low a speed.

It may be seen that for our stocks there is a significant drop in tensile with increasing speed up to about 10 percent/msec. Somewhere near this speed (before it if not at it) the tensile goes through a minimum and rises more or less drastically until at the highest speeds observed it is, for the tread stocks, 100 percent (or more) greater than the Scott tensile. The Hevea gum tensile at highest speed is about the same as the Scott tensile, but the tensile of the GR-S gum is multiplied sevenfold.

Effect of Curing

Figure 5(a) shows ultra speed tensiles obtained on raw Hevea gum and tread stocks. The samples for these runs were interspersed in random order among those of the replication series. It may be seen that raw smoked sheet, which has zero Scott tensile, attains a tensile of 900 lb./in.^{2} at highest speed. Compounded (but uncured) tread stock has a Scott tensile of 530 lb./in.^{2}. This increases to 4200 lb./in.^{2} at highest ultra speed.

Dependence on Elastomer

Tensile data on the elastomers of the second series are to be found plotted in Figs. 6(a)–11(a). In this series GR-S does not show a minimum in tensile (it shows a slight minimum in the Replication series), but a steady rise as speed is increased. Hycar OR and Perbunan also show a steady rise. The other synthetics show the minimum as Hevea does.

It is to be noted that the stocks which exhibit the minimum (except for GR-S) are ones which are known to crystallize at ordinary slow rates of stretch. The other stocks do not crystallize. The minimum in the first GR-S series may be related to an orientation, of lower order than crystallization.

Dependence on Temperature

Figures 6–11(a) also show the results of a study of effect of temperature on ultra speed tensile. In general,

ultra speed tensile of gum stocks at 240 percent/msec. is reduced by 67 to 75 percent by heating to 150°C; of tread stocks, it is reduced only by 20 to 50 percent (exception: possibly Perbunan). At 150°C, gum GR-S has such a poor ultra speed tensile that the pin on the wheel knocks out from the sample by a clean cut an exact replica of the pin.

ELONGATION AT BREAK

Dependence on Speed

Figure 4(b) gives the elongation at break obtained from measurement of the replication series. It shows at 1 percent/msec. elongation values of Roth and Holt. The Hevea stocks, especially the gum, show a maximum in elongation at break twice that at Scott speed, between 10 and 30 percent/msec. The GR-S stocks do not exhibit this phenomenon to as marked an extent, if at all. At highest ultra speed all stocks become drastically shortened.

Effect of Curing

See Fig. 5(b). Elongation at break of the uncured smoked sheet and tread compound shows curves similar to those of the cured stocks, except for an upward displacement of the tread stock. The ultra speed values of the uncured smoked sheet agree fairly closely with those at the same speeds for the cured gum stocks.

Dependence on Elastomer

See Figs. 6(b)-11(b). The different elastomers, in general, are classifiable into two groups; one in which the elongation at break passes through a maximum at an intermediate ultra speed (such as Hevea, and Butyl) and one for which it steadily decreases with increasing speed of break. Except for Neoprene GN, this classification is the same as that of elastomers which do or do not crystallize on stretching.

Dependence on Temperature

Ultra speed elongation at break at 150° seems, in general, to be greater for tread stocks than for gum stocks at all speeds except the highest. Heating to 150° *reduced* elongation of all stocks except butyl tread.

Modulus

The ultra speed modulus was computed from the tensile and elongation at break on the approximation that the ultra speed stress-strain curve is linear. For Scott speed the modulus was arbitrarily taken as the slope of the line between the origin of the (single-pull) autographic stress-strain curve and the point of inflection. This value thus does not include the subsequent increase in modulus which many elastomer stocks undergo at higher elongations (usually ascribed to orientation). It lies between the Wall-Young's modulus and the shearing modulus as computed from Young's modulus by dividing the latter by three. Within this range, the comparability of the low speed values with the ultra speed values is somewhat uncertain.

Dependence on Speed

Figure 4(c) gives modulus measurements obtained from the Replication series. It contains at 1 percent/msec. modulus values of Roth and Holt. Modulus is almost independent of elongation rate up to 10 percent/msec., above which speed it rises rapidly.

Effect of Curing

Modulus of uncured smoked sheet and of uncured tread compound shows curves similar to those of the

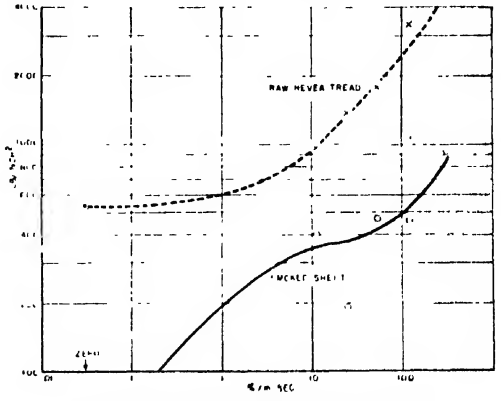


FIG. 5(a). Tensile—unvulcanized Hevea.

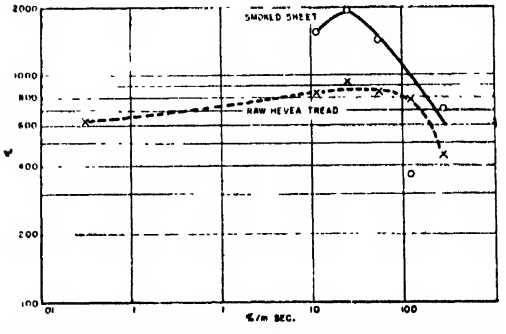


FIG. 5(b). Elongation at break—unvulcanized Hevea.

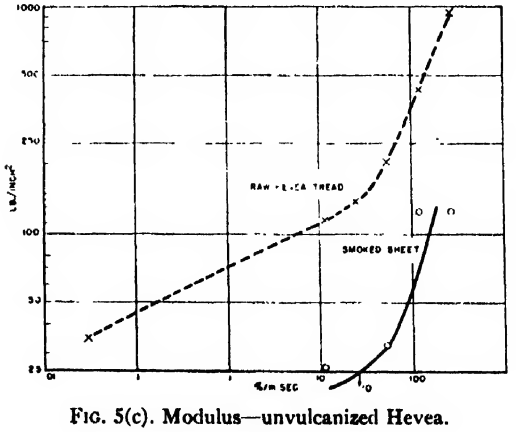


FIG. 5(c). Modulus—unvulcanized Hevea.

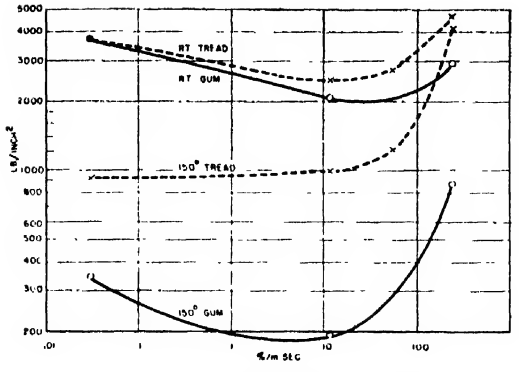


FIG. 6(a). Tensile—Hevea ADJ, K.

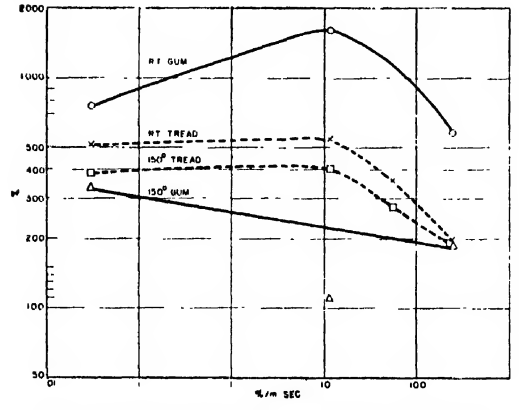


FIG. 6(b). Elongation at break—Hevea ADJ, K.

cured stocks, except for a downward displacement in value. See Fig. 5(c).

Dependence on Elastomer

Plots of moduli of the elastomer series, in general, show that modulus is more or less constant until the ultra speeds are reached, when it starts to increase with an accelerated pace.

Dependence on Temperature

Modulus of gum stocks is little affected by increasing temperature to 150°; that for tread stock is shifted to the side of higher elongation rate. Exceptions to this last generalization are Hevea and Neoprene, where the curves for tread stocks rise to the same modulus at highest speed.

Energy of Break

Ultra speed energy of break was computed from the tensile and elongation at break on the approximation that the ultra speed stress-strain curve is linear. For Scott speed, energy of break was obtained by integrating (by planimeter) the complete single-pull autographic stress-strain curve after extending it to the Scott tensile and elongation at break. The slow speed energy thus includes the contribution of all orientation effects. Energy of break values appear to be subject to an excessively large error. They fluctuate so greatly that no trends with speed appeared significant.

Independent Determination of Energy of Break

Energy of break was checked independently by measuring the instantaneous drop in wheel speed occurring when the sample is broken. This was measured by connecting the oscillograph to give a changing pattern, the frequency of which was the difference between the wheel frequency and an adjustable reference frequency. The difference between this difference frequency before and immediately after breaking a sample was combined with the moment of inertia of the wheel to compute the energy of break.

Two values obtained are 18.5 and 15.6 ft. lb. for rubber tread. These values, determined at 25 cycles/sec. wheel speed, are to be compared with the total energy of break, 10.12 ft. lb., given by a stress-strain curve at 23 cycles, of a similar sample. The latter figures include 1.69 ft. lb. kinetic energy of the end thrown off. The high tensiles indicated by our measurements of the stress-strain curves are thus independently confirmed.

DISCUSSION

Crystallization and Relaxation

In explaining the observed effects of elongation rate on the stress-strain properties of elastomers, it is necessary to consider two phenomena which can be induced in an elastomer by stretching, but which require finite

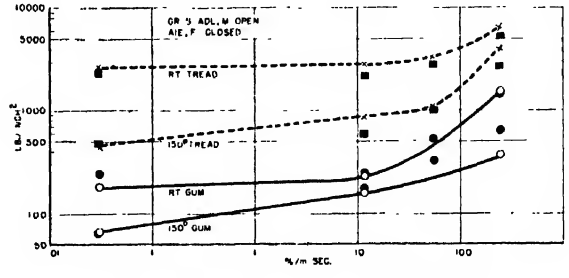


FIG. 7(a). Tensile—GRS. ADL, M; AIE, F.

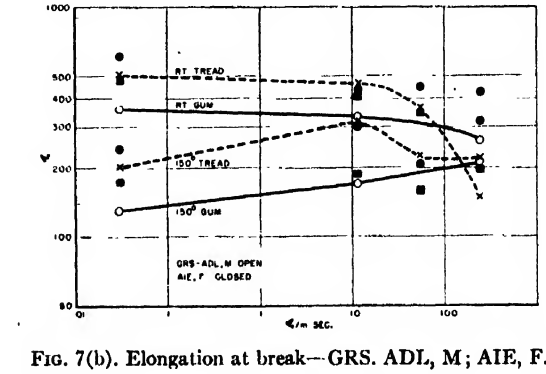


FIG. 7(b). Elongation at break—GRS. ADL, M; AIE, F.

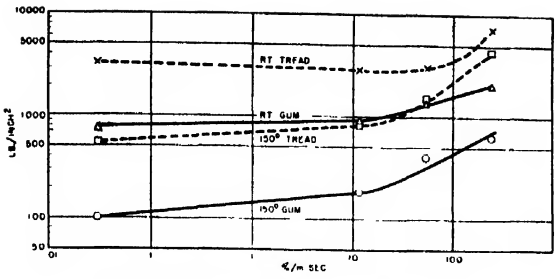


FIG. 8(a). Tensile—Hycar OR. ADN, O.

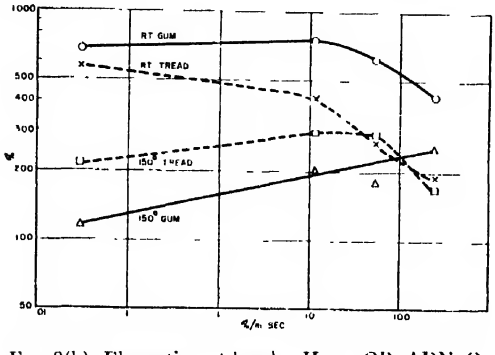


FIG. 8(b). Elongation at break Hycar OR, ADN, O

time for completion. These phenomena are, first, crystallization and, second, relaxation.

The fact that rubber crystallizes on stretching is familiar to all rubber chemists. Many, if not most, attribute the upbend in the ordinary stress-strain curve to onset of crystallization. According to Acken, Singer, and Davey,[8] a minimum time of the order of 1.2 sec. is required after stretching before enough crystallization develops to show up in a x-ray picture. As our ultra speed curves are produced in times of lower order of magnitude than this, it is plausible to attribute differences between those obtained at the lesser ultra speeds and those obtained at slow speed to lack of time for crystallization to develop. This lack of time would account for the approximate linearity of the ultra speed stress-strain curves.

This theory may be used to account for the differences in tensile and elongation at break between the slow autographic stress-strain and the mean ultra speed curve of Fig. 3. Since the latter was produced in an average time of 28 msec., presumably much too short for crystallization, we may conclude that crystallization increases tensile strength but shortens the elongation at break.

A corollary of the hypothesis of a finite time requirement for completion of crystallization is of interest to problems of abrasion resistance. Inasmuch as abrasion processes generally occur in cycles shorter than 20 msec., it appears questionable whether the crystallization characteristics of a tread stock can have any direct effect on its abrasion resistance.

Relaxation, the second phenomenon invoked to explain the ultra speed results, is the loss of stress with time in a sample held at a constant deformation. It is thought to be due to a slow intermolecular slippage within the rubber, accompanied sometimes by a reversible, sometimes irreversible, breaking of bonds. Kuhn[9] and others have suggested that many properties of rubber may be explained by postulating more than one type of slippage bonds, the different types being distinguished by different relaxation rate constants.

Now, if the time of test is made shorter than the relaxation time of the rapidly relaxing components they contribute more and more toward the total modulus. This is how we account for the increase in modulus as rate of elongation increases to 200 percent/msec. and greater. The increase in tensile with increasing speed implies that the slipping bonds also add to strength when not given time to slip.

The minima in tensile and maxima in elongation observed with crystallizing stocks now may be interpreted as the point of balance between decreasing importance of crystallization and increasing importance of slipping bonds that fail to slip as speed is increased. If the 90 percent drop in tensile of gum Hevea which is observed at Scott speed on heating from room temperature to 150° is a measure of the contribution of crystallinity, the formation of only about one-third of the total crystallinity is prevented by stretching at the lower ultra speeds.

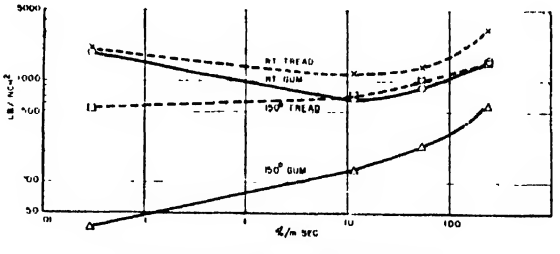


FIG. 9(a). Tensile—butyl. ADP, Q.

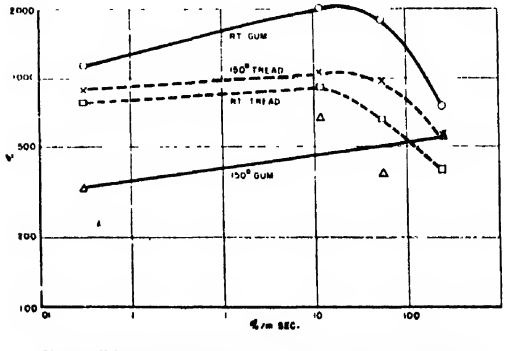


FIG. 9(b). Elongation at break—Butyl. ADP, Q.

[8] Acken, Singer, and Davey, Ind. Eng. Chem. 24, 54–7 (1932), R.C.T. 5, 30–8 (1932).

[9] W. Kuhn, Zeits. f. physik Chemie 42B, 1–38 (1939).

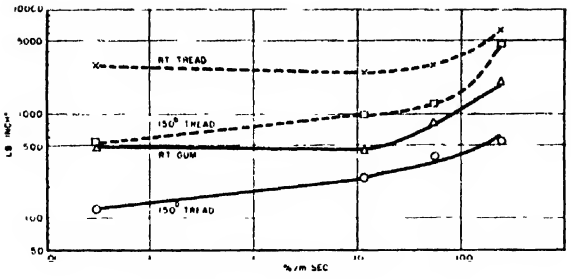


FIG. 10(a). Tensile—Perbunan, ADR, S.

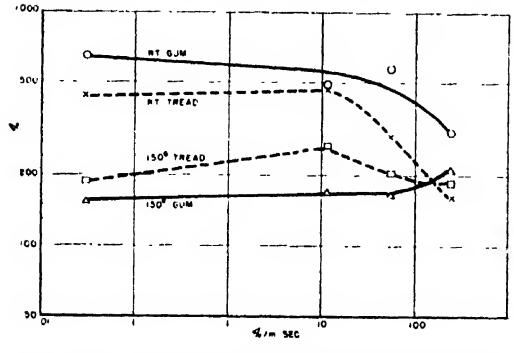


FIG. 10(b). Elongation at break—Perbunan, ADR, S.

On the slipping bond hypothesis the elongation rate at which the sharp rise occurs, 200 percent/msec., qualitatively gives the order of magnitude of the relaxation time as 0.5 msec. It is believed that many service processes may occur within times of this order of magnitude. For example, Markwick and Starks[10] report high frequency oscillations of 1400 cycles/sec. in tread rubber retracting from a dry road. As far as we know, no one has reported confirmation of this observation and it may possibly be a spurious mechanical effect. However, if an important part of the strength of a stock originates from rapidly relaxing elements, it is clear that a completely successful correlation between laboratory and service can never be expected until the time of test is reduced to that characteristic of the service condition.

Heat of Activation of Slippage

If the above hypothesis be true that the rise in modulus and tensile may be attributed to bonds, which ordinarily completely relax within the time of stretching, not having time to relax completely at the higher ultra speeds, then it would be reasonable to expect to find a particular curve to shift bodily to the side of higher elongation rate on going to higher temperature. For, if one used as abscissae the parameter, speed/relaxation rate (or speed×relaxation time), increase of temperature increases relaxation rate, so to find the same modulus one should look for it at a greater speed. A complete shift of the curve without change in shape is too much to expect, inasmuch as it could occur only in the unlikely event of only one relaxation type. With more than one type, the modulus contribution of each would be shifted by a different amount corresponding to its different change in relaxation rate. This would bring about a change in shape of the composite curve which is the sum of the individual components.

We have noted the above shift for the case of carbon black stocks. A heat of activation may therefore be roughly computed from the ratio of rates at 150° and at room temperature. A list of such values is given in Table IV.

It may be seen that the heat of activation is of the order of magnitude of 3 kg cal. This may be compared with 10.3 kg cal. heat of adsorption of benzene on graphite[11] (the nearest similar process to adsorption of rubber on carbon black on which there are published data), 10 kg cal. heat of activation of shear in raw rubber,[12] 15 kg cal. for shear in raw GR-S,[13] 15 kg cal. for transient relaxation in cured rubber, and 15 kg cal. for steady relaxation in cured rubber.[14] The considerably lower value of 3 kg cal. is in accord with a "slide-fastener" mechanism of slippage of rubber molecules along a carbon black surface proposed by the present author. In this mechanism the attached segments of rubber are supposed to slip with moderate resistance along the surface of the black from one end to the other

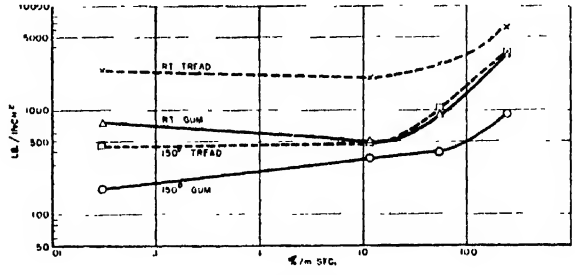


FIG. 11(a). Tensile—Neoprene GN. ADT, U.

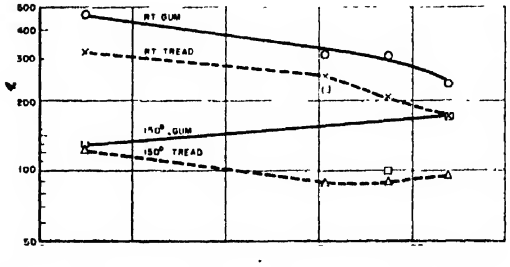


FIG. 11(b). Elongation at break—Neoprene GN. ADT, U.

[10] (a) Report of Road Research Board for the year ending March 31, 1939, pp. 150–6, Great Britain Department of Science and Industry Research. (b) A. D. H. Markwick and H. J. H. Starks, J. Inst. Civil Eng. (London) **16**, 309–25 (1941).

[11] W. D. Harkins and G. E. Boyd, J. Am. Chem. Soc. **64**, 1195–1204 (1942), Table III.

[12] H. M. Smallwood, J. App. Phys. **8**, 505 (1937).

[13] M. Mooney (private communication).

[14] Mooney, Wolstenholme, and Villars, J. App. Phys. **15**, 324–337 (1944).

without pulling away free from it. When the rubber-black bond reaches the end of the "zipper" it is fairly permanently resistant to complete separation.

The failure to find a shift in the gum stock curves means, on the above hypothesis, that the heat of activation of slippage for the other types of slippage is very low; i.e., the relaxation rates involved have too low a temperature coefficient to be measurable within the accuracy of the experiment.

TABLE IV. Heat of activation of relaxation rate in tread stocks.

Elastomer	ΔH, kg cal.
Hevea	3.08
GR-S	3.67
Hycar OR	3.75
Butyl	3.77
Perbunan	2.78
GR-M	1.57

ACKNOWLEDGMENT

Appreciation is herewith expressed to the various members of the staff who have been consulted in the development of the machine; in particular to Drs. E. M. Grabbe, M. Mooney, and S. A. Black. Mr. R. B. Frost designed and supervised the original construction of the machine and Mr. T. J. Rhodes designed several of the later modifications. The experimental work of operating the machine was carried out by Mr. O. K. Laible.

Relations Concerning Wave Fronts and Reflectors

K. S. KELLEHER
Naval Research Laboratory, Washington, D. C.

(Received October 20, 1949)

By means of a vector notation for surfaces, relations are derived among an incident wave front, reflector and reflected wave front. A method is introduced for evaluating the deviation of a wave front surface from a plane. Problems are included in order to indicate the simplicity and utility of the analysis.

INTRODUCTION

IN working with microwave antennas, it is desirable to know the characteristics of their wave fronts. Some knowledge can be obtained by employing geometrical optics principles to trace a few selected rays. It is the intention of the present paper to consider the reflection of a normal congruence of rays, defined by a wave front surface.[1] The problem presented and solved is: Given any two of the three surfaces, incident wave front, reflector and reflected wave front, find the third.

THE REFLECTED WAVE FRONT

The incident wave front surface $X(u, v)$[2] and the reflector $R(s, t)$ are indicated in Fig. 1. The reflected wave front $Y(u, v)$ can be obtained by using the definition that all points of $Y(u, v)$ lie at equal ray path length from $X(u, v)$ together with the fact that rays are orthogonal to the wave front surfaces. From the figure, two values of the general reflected ray can be obtained and equated.

$$Y-R=(C-|R-X|)\xi, \tag{1}$$

where C is the total ray path length from X to Y; $|R-X|$ is the length of the incident ray; and ξ is a unit normal to the reflected wave front, Y.

The vector, ξ, is determined[3] from the fact that incident and reflected angles at the reflector surface are equal.

$$\xi=r-2\eta(\eta\cdot r) \tag{2}$$

where $r=(R-X)/|R-X|$ and η is the unit normal to the reflector at the point of incidence.

Placing (2) in (1) and reducing, there results

$$Y=C\xi+X+2\eta[\eta\cdot(R-X)]. \tag{3}$$

For most practical purposes, it is convenient to consider the reflected wave front surface at zero path length ($C=0$) from the incident wave front. This fictitious wave front is a parallel surface[4] to any actual reflected wave front and therefore can be used to define the reflected congruence of rays. Such a wave front could be subjected to further reflections in the same way that X was reflected from R.

Substituting $C=0$ in (3), the reflected wave front is

$$Y(u, v, s, t)=X(u, v)+2\eta(s, t)[\eta(s, t)\cdot\{R(s, t)-X(u, v)\}]. \tag{3a}$$

The parentheses are used to indicate the dependence of the functions on the various parameters involved. A relation between u, v and s, t is required in order that Y should be a function of only two parameters. This is obtained from the condition that the incident rays are normal to the incident wave front. Using tangent vectors to the surface, this condition is written

$$(R-X)\cdot X_u=0 \quad \text{and} \quad (R-X)\cdot X_v=0, \tag{4}$$

[1] W. Blaschke, *Differential Geometric I* (Dover Publications, New York, 1945), p. 104.
[2] This vector notation is discussed in Blaschke.
[3] L. Silberstein, *Simplified Method of Tracing Rays Through Any Optical System* (Longmans, Green and Company, London, 1918), p. 1.
[4] L. P. Eisenhart, *An Introduction to Differential Geometry* (Princeton University Press, Princeton, 1940), p. 272.

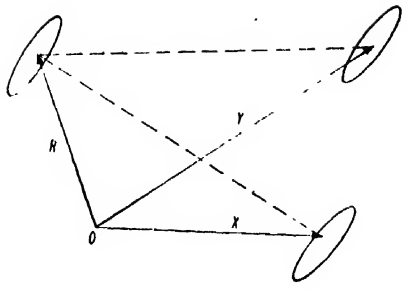


FIG. 1.

where the subscripts denote partial differentiation. From these expressions $s=s(u, v)$ and $t=t(u, v)$ can be obtained and when they are substituted in (3a) yield the desired reflected wave front surface $Y(u, v)$.

Since the incident and reflected wave fronts could be interchanged without altering the above treatment, there remains only the problem of determining the reflector which transforms a given incident wave front into a given reflected wave front.

THE REFLECTOR

In order to find this reflector, use (1) with the sign of the unit normal reversed, so that ξ points toward the reflector.

$$R=Y+(C-|R-X|)\xi. \tag{5}$$

As before, ξ is the normal to the wave front $Y(p, q)$, and $X=X(u, v)$ is the second wave front surface. In this expression, $|R-X|$ can be evaluated by subtracting X from both sides and taking the scalar product of each side with itself. The result reduces to

$$|R-X|=\frac{C^2+(Y-X)^2+2C(Y-X)\cdot\xi}{2[C+(Y-X)\cdot\xi]}.$$

Using this, (5) becomes

$$R(u, v, p, q)=Y+\xi\frac{[C^2-(Y-X)^2]}{2[C+(Y-X)\cdot\xi]}=Y+\xi G. \tag{5a}$$

R is obtained as a function of u and v alone through use of the condition that $(R-X)$ is normal to the surface X.

$$(R-X)\cdot X_u=0 \quad \text{and} \quad (R-X)\cdot X_v=0,$$

or

$$(Y-X+\xi G)\cdot X_u=0 \quad \text{and} \quad (Y-X+\xi G)\cdot X_v=0. \tag{6}$$

These two equations are solved for $p(u, v)$ and $q(u, v)$ which when substituted in (5a) gives R as a function of u and v alone.

DEVIATION FROM A PLANE WAVE

Given a wave front surface, it is normally desirable to know its deviation from a plane, since such knowledge gives an indication of the far field radiation pattern. Referring to Fig. 2, let the wave front be represented as $Y(u, v)$ and designate a unit normal to the plane as N. Introduce a constant vector Y_0 from the origin to a point of tangency of the plane with the surface. From the figure, it is evident that the deviation, Δ, is the projection of the vector $(Y-Y_0)$ on N. Analytically stated,

$$\Delta(u, v)=|Y-Y_0|\cos\beta=(Y-Y_0)\cdot N. \tag{7}$$

Therefore, for any point (u, v) of the surface, the deviation can be determined.

CONICAL WAVE FRONT AND PARABOLIC CYLINDER

A line source with linear phase distribution produces a conical wave front.[5] If the source is positioned along the

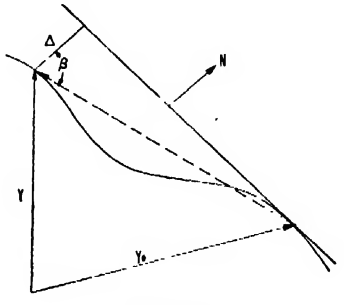


FIG. 2.

focal line of the reflector (Fig. 3), the incident wave front may be written

$$X=Fi+uj+vk$$

where

$$F=[(h-v)^2\tan^2\alpha-(u-f)^2]^{\frac{1}{2}}.$$

The reflector is

$$R=si+(s^2/4f)j+tk.$$

The normal to the surface R can be obtained from the vector product of two tangent vectors.

$$R_s\times R_t=[i+(s/2f)j]\times k.$$

The unit normal then is

$$\eta=[si-2fj][s^2+4f^2]^{-\frac{1}{2}}.$$

Upon substitution of these values in (3a), the reflected wave front becomes

$$Y=i\frac{[s(s^2+4uf)-(s^2-4f^2)F]}{s^2+4f^2}+j\frac{[s^2(u-2f)-4uf^2+4fsF]}{s^2+4f^2}+kv.$$

This expression for Y involves three parameters, so that a given point on the incident wave front $X(u, v)$ does not yield a single value of Y, but rather a value which depends on s. It is therefore necessary to find an auxiliary

[5] K. S. Kelleher, "Antenna Wavefront Problems," Naval Research Laboratory Report R-3530.

relation between s, u and v. This relation, obtained from the condition $(R-X)\cdot X_u=0$ is $(s^2-4f)F=4fs(u-f)$. Substituting this and reducing, Y becomes

$$Y=si+[(f-u)\sec\theta-f]j+vk$$

where

$$\sec\theta=(f+s^2/4f)/(f-s^2/4f).$$

From Fig. 4, the angle θ can be related to the cone angle α.

$$(f-u)\sec\theta=(h-v)\tan\alpha.$$

Thus the reflected wave front can also be written

$$Y=si-[(f-h\tan\alpha)+v\tan\alpha]j+vk.$$

This surface is a plane, intersecting the $x-y$ plane in the line $y=-(f-h\tan\alpha)$ and at an angle $\pi-\alpha$. (This can be verified by using (7) to show that the deviation of this Y surface from such a plane is zero.)

POINT SOURCE AND SPHERICAL REFLECTOR

When a point source is used to feed a spherical reflector, it is possible to obtain a nearly plane wave front characteristic by placing the feed near the half-radius point of the reflector.[6] The positioning of the feed as well as the deviation of the wave front from a plane can be found using (3a). Choose a spherical reflector with center at the origin. For convenience, normalize all points to the radius of the sphere. Then the reflector, unit normal and feed vectors may be written:

$$R=xi+(1-r^2)^{\frac{1}{2}}j+zk,$$
$$\eta=R,$$
$$X=jd,$$

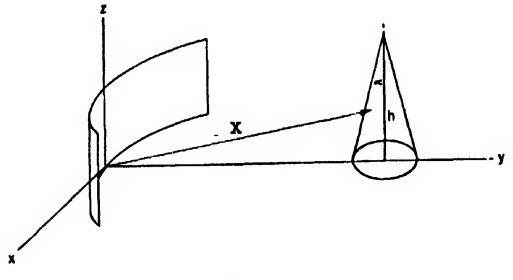


FIG. 3.

where

$$r^2=x^2+z^2.$$

The wave front, Y, is found by substituting these vectors in (3a)

$$Y=Dxi+[D(1-r^2)^{\frac{1}{2}}+d]j+Dzk,$$

where

$$D=2-2d(1-r^2)^{\frac{1}{2}}.$$

The deviation of this wave front from a plane is found by using (7) with $N=j$ and $Y_0=(2-d)j$

$$\Delta=2(1-r^2)^{\frac{1}{2}}+2(dr^2-1).$$

It is evident that $\Delta(0)=0$, but this deviation will also vanish at another value of r if d, the feed position, is properly chosen. Setting $\Delta=0$ and solving for d, there results,

$$d=[1-(1-r^2)^{\frac{1}{2}}]/r^2.$$

Therefore, given a value of r for which the deviation, Δ, should vanish, the required feed position, d, can be immediately determined.

Once the value of d has been established, it is possible to find the maximum value of $\Delta(r)$ in the usual manner. The result is

$$\Delta_{max}=2(d-0.5)^2/d$$

for a reflector of unit radius.

REFLECTORS WHICH YIELD VIRTUAL POINT SOURCE

Occasionally in microwave antenna design it is advisable to use a virtual point source feed.[7] A catalog of the reflectors which will produce a spherical wave front from a point source feed, can be obtained from Eq. (5a). From the analysis, the reflector desired is

$$R=Y+\frac{C^2-(Y-X)^2}{2[C+(Y-X)\cdot\xi]}\xi \qquad (5a)$$

where Y and X in this case are points. Let $X=0$ and $Y=ai$; the unit normal to the Y wave front is

$$\xi=(x-a)i+yj+zk$$

where

$$y=[1-z^2-(x-a)^2]^{\frac{1}{2}}.$$

Substituting these values, the reflector becomes

$$R=\frac{[(x-a)(C^2+a^2)+2aC]i+y(C^2-a^2)j+z(C^2-a^2)k}{2[C+a(x-a)]}.$$

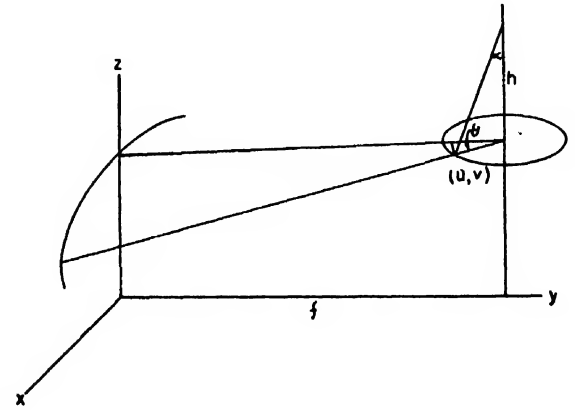


FIG. 4.

[6] J. Ashmead and A. B. Pippard, "The Use of Spherical Reflectors as Microwave Scanning Aerials," J. Inst. Elec. Eng. 93, 627 (1946).

[7] R. F. Rinehart, J. App. Phys. 19, 862 (1948).

In terms of components, r, s and t along the x, y and z axes,

$$R=ri+sj+tk.$$

It is possible to eliminate the parameters x, y and z and obtain

$$\frac{(r-a/2)^2}{C^2/4}+\frac{s^2+t^2}{(C^2-a^2)/4}=1.$$

From inspection of this equation, the following tabulation may be made:

For $C>|a|$, the reflector is an ellipsoid, unless $a=0$ in which case it is a sphere of radius $C/2$. For $C<|a|$, the reflector is a hyperboloid of revolution, unless $C=0$ in which case it is the plane, $x=a/2$. For $C=a$, the result is $r=0$ or $r=a$, and so has no practical meaning.

REFLECTOR WHICH WILL PRODUCE A PLANE WAVE

A less trivial application of (5a) is had when it is desired to find the reflector which will transform an arbitrary incident wave front $X(u, v)$ into a plane wave front $Y(p, q)$.

If the general wave front surface is given by $y=y(x, z)$, in vector notation it can be written

$$X=ui+F(u, v)j+vk.$$

Let the plane wave front be the $x-z$ plane so that

$$Y=pi+qk$$

and the unit normal to this wave front, $\xi=j$. When these values are substituted into (5a), there results

$$R=Y+G\xi=pi+Gj+qk. \tag{8}$$

Here

$$G=\frac{C^2-(Y-X)^2}{2[C+(Y-X)\cdot\xi]}=\frac{C^2-(p-u)^2-F^2-(q-v)^2}{2[C-F]} \tag{9}$$

where C is the constant path length from wave front X to wave front Y.

In order to obtain R as a function of only two parameters, use (6) which relates the four parameters u, v, p, q.

$$(R-X)\cdot X_u=(p-u)+(G-F)F_u=0,$$

and

$$(R-X)\cdot X_v=(q-v)+(G-F)F_v=0. \tag{10}$$

The substitution of values from (10) into (9) yields

$$2G(C-F)=C^2-F^2-(G-F)^2(F_u^2+F_v^2).$$

Subtract $2F(C-F)$ from both sides and the equation becomes a quadratic in $(G-F)$,

$$2(G-F)(C-F)=(C-F)^2-(G-F)^2(F_u^2+F_v^2)$$

which has the solution

$$G-F=(F-C)[1\pm(1+F_u^2+F_v^2)^{\frac{1}{2}}][F_u^2+F_v^2]^{-1}. \tag{11}$$

The ambiguity of sign can be resolved in a particular case.

Since $G=G(u, v)$ can be obtained from this equation, (10) and (11) yield $p=p(u, v)$ and $q=q(u, v)$. Therefore the reflector R in (8) is given as a function of u and v alone. In some simple problems, it is possible to eliminate the parameters u and v to obtain the reflector surface as $G=G(p, q)$.

As an example of the preceding analysis, let the wave front $X(u, v)$ be that produced by a source along the z axis whose phase variation is such as to yield the wave front surface $y^2=[(r^2-z^2)^{\frac{1}{2}}-a]^2-x^2$. In this case $y=F$, $x=u$, $z=v$.

In order to use (11), it is first determined that

$$y_x^2+y_z^2=(x^2+z^2T^2)/y^2$$

where

$$T^2=(x^2+y^2)/(r^2-z^2).$$

Then, from (11) one component of R is

$$G=y+\frac{y(y-C)[y\pm rT]}{x^2+z^2T^2}=y(1+D).$$

When this value is substituted into (10) the other two components of R become

$$p=x(1+D),$$
$$q=z(1+DT).$$

Points on the reflector surface are obtained by computing G, p, and q for appropriate values of x, y, z. If z is set equal to zero, the reflecting curve is one of two parabolas, due to the ambiguity of sign. The correct reflector corresponds to the negative sign of the radical in (11).

As a check of the method, it might be noted that setting $a=0$ yields a spherical wave front, such that $T=1$ and $D=(C-y)/(r+y)$. Upon elimination of y from the components of R, there results

$$p^2+q^2=-(C+r)[2G-(C+r)]$$

which is the expected paraboloid surface.

The Effect of an Obstacle in the Fresnel Field on the Distant Field of a Linear Radiator

G. A. WOONTON
Department of Physics, McGill University, Montreal, Quebec, Canada
(Received October 19, 1949)

At radio wave-lengths the use of mirrors and lenses in the measurement of radiators, by methods which are customary in physical optics, is possible only if errors, due to incidental diffraction from the relatively small apertures of such apparatus, can be corrected by calculation. Theory is developed in this paper from which these calculations can be derived for linear radiators. A new relation which expresses the Fresnel field in terms of the distant field is fundamental to the method. Satisfactory agreement has been obtained between the predicted and measured distant fields of several electromagnetic horns whose near fields were obstructed; the dispersion between measured and predicted values is given as under one-half decibel but the details of the experiment are not reported.

INTRODUCTION

At radio wave-lengths, the information from optical measurements often is obscured by incidental diffraction from the apertures of the component lenses, mirrors, and slits of the optical system. Incidental diffraction, although present in optical systems for visible light, can usually be neglected, but in the radio case, even for the shortest microwaves, the ratio of aperture dimension to wave-lengths is such that these effects, always, must be considered; practical limitations on physical dimensions and weight make it impossible to scale radio-optical devices in proportion to wave-length so that the ratio of aperture diameter to wave-length in the radio case is different by a factor of at least 10^{-5} from the ratio used in apparatus for measurements with visible light. It is for this reason that the application of optical techniques to radio measurements depends on the ability of the experimenter to calculate instrument errors and to use the calculations to plan the location of apertures so as to minimize the disturbance that they cause.

A problem of this nature, which is fundamental to many radio measurements, can be stated in specific terms; the radiation field from a linear radiator, for example a sectoral electromagnetic horn, is obstructed by a second aperture placed a short distance in front of it and at an arbitrary angle to its normal;[1] find the distant field produced by the combination. The analysis which follows is a general solution of this problem, and is capable of quantitative application to particular systems.

OPTICAL THEORY

An experimental investigation reported by Woonton, Hay, and Vogan[2] has shown that Kirchhoff's scalar formulation of Huygen's principle[3] is adequate to describe the distant diffraction field of a rectangular aperture whose dimensions are as small as 10 wave-lengths to a side and to describe the near field of the same aperture at a distance no greater than 10 or 20 wave-lengths from its opening. In the same investigation it was verified, also, that as in the optics of visible light, the complete fields can be calculated as the product of two independent, one-dimensional fields due separately to the electric distribution along the X axis of the aperture and along the Y axis. For these reasons, the solution which follows is based on the Kirchhoff formula and assumptions, and expresses the fields in terms of one coordinate only.

A system of Cartesian coordinates is chosen so that the origin lies at the intersection of the diagonals of a plane, rectangular, radiating aperture which will be called S_1; for first consideration S_1 is not rotated and lies wholly in the XY plane. The distribution of electric intensity along its X axis is given in both phase and amplitude by $F_1(x)$. At a distant point in the XZ plane, $P_1(R_1\theta_1)$ (where R_1 is the radius vector from the origin to P_1 and θ_1 is the angle between R_1 and the Z axis), the distant field (subject to the usual approximations, which include the requirements that θ_1 is not too great) is:

$$\varphi_1\left(R_1, \frac{\sin\theta_1}{\lambda}\right) = Re\frac{(\cos\theta_1+1)}{R_1}\exp(j[\omega t - kR_1])$$

$$\times Y\left(\frac{j}{2\lambda}\right)^{\frac{1}{2}}\int_{-\infty}^{\infty} F_1(x)\exp(-j2\pi x\gamma_1)dx, \quad (1)$$

in which λ = wave-length of the radiation, $k=2\pi/\lambda$, ω = angular velocity of the field, $j=(-1)^{\frac{1}{2}}$, Y = a factor which remains constant as long as P is confined to the XZ plane and which is an integral of the same form as that which appears in (1) but is related to aperture distribution in the Y direction, $\gamma_1=-(\sin\theta_1)/\lambda$.

Equation (1) represents a circular wave, expanding from the origin, whose front at radius R_1, in the XZ plane, is modified in phase and amplitude by the integral factor:

$$G_2(\gamma_1)=\int_{-\infty}^{\infty} F_1(x)\exp(-j2\pi x\gamma_1)dx. \quad (2)$$

[1] Woonton, Borts, and Carruthers, J. App. Phys. **21**, 428 (1950).
[2] Woonton, Hay, and Vogan, J. App. Phys. **20**, 71 (1949).
[3] J. S. Slater and N. H. Frank, *Electromagnetism* (McGraw-Hill Book Company, Inc., New York, 1947), p. 167.

(The use of the negative sign in the definition will be clear later.)

The distribution $F_1(x)$ produces a Fresnel distribution of electric intensity in front of the radiator, as well as the field at a great distance. At $z=b$ define an auxiliary set of (ξ, η) coordinates in which the ξ- and η-axes are parallel to those of X and Y and take the new origin at $(0, 0, b)$. The distribution produced along the ξ-axis by $F_1(x)$ is $F_2(b, \xi)$, which for convenience will be described as $F_2(\xi)$. At a distant point $P_2(R_2\theta_2)$ (where R_2 is the radius vector from $z=b$ to P_2, and θ_2 is measured at $(0, 0, b)$ the distribution $F_2(\xi)$ may be thought of as producing the field instead of $F_1(x)$. This field is given by:

$$\varphi_2\left(b, R_2, \frac{\sin\theta_2}{\lambda}\right) = Re\frac{(\cos\theta_2+1)}{R_2}\exp(j[\omega t - kR_2])$$

$$\times Y''\left(\frac{j}{2\lambda}\right)^{\frac{1}{2}}\int_{-\infty}^{\infty} F_2(\xi)\exp(-j2\pi\xi\gamma_2)d\xi \quad (3)$$

where

$$\gamma_2 = -(\sin\theta_2)/\lambda.$$

Equation (3) describes a circular wave expanding from $(0, 0, b)$ whose front at radius R_2 (measured from $z=b$) is modified in amplitude and phase by the integral factor

$$G_2(\gamma_2) = \int_{-\infty}^{\infty} F_2(\xi)\exp(-j2\pi\xi\gamma_2)d\xi. \quad (4)$$

The relation between the integral factors (2) and (4) which appear in Eqs. (1) and (3) can be found by comparison of (1) and (3) after (1) has been modified to express the field along the circular wave front marked out by P_2 as θ_2 varies. Let $R_1=(R_2+b)$ so that Eqs. (1) and (3) express the same field at $\theta_1=\theta_2=0$. Choose R_1 and R_2 both very large compared to b, so that in the amplitude factors of the two equations R_1 can be interchanged with R_2 and so that a single angle θ can be used to represent the θ-coordinate of corresponding points on the two circular wave fronts instead of separate angles θ_1 and θ_2. Because Eqs. (1) and (3) both express distant fields, one can reason that at any angle θ the two equations differ only in phase and this difference can be traced to two causes: the translation of the origin directly and the separation of the perimeters of the two circles which the translation of the origin causes. The first phase difference is constant and can be written by substitution:

$$\exp(-jkR_1) = \exp(-jk[R_2+b]) = \exp(-jkR_2)\exp(-jkb). \quad (5)$$

The second phase difference arises because Eq. (1) describes the field along a circle of radius (R_2+b) with center at the origin while Eq. (3) describes the field along a circle of radius R_2 with center at $(0, 0, b)$. At any angle, θ, the perimeters of the two circles are separated by a distance, δ, which can be written to an approximation that improves with the magnitude of R_1 and R_2 as:

$$\delta \doteqdot b(1-\cos\theta) \doteqdot b(\sin^2\theta)/2. \quad (6)$$

Equation (1) will describe the field along the circle marked out by R_2 only if its phase is corrected by the constant factor (5) and advanced by a variable factor proportional to the distance, δ, namely:

$$\exp(jkb\sin^2\theta/2) = \exp(j\pi b\lambda[\sin\theta/\lambda]^2). \quad (7)$$

Equate the modified Eq. (1) to Eq. (3) and cancel the common factors:

$$Y\exp(-jkb)\exp\left(j\pi b\lambda\left[\frac{\sin\theta}{\lambda}\right]^2\right) \times \int_{-\infty}^{\infty} F_1(x)\exp(-j2\pi x\gamma)dx = Y'\int_{-\infty}^{\infty} F_2(\xi)\exp(-j2\pi\xi\gamma)d\xi. \quad (8)$$

Define

$$G_3(\gamma) = \exp\left(j\pi b\lambda\left[\frac{(\sin\theta)}{\lambda}\right]^2\right);$$

then in the abbreviated notation:

$$Y\exp(-jkb)G_3(\gamma)G_1(\gamma) = Y'G_2(\gamma). \quad (9)$$

Because relative, rather than absolute amplitudes and phases are important in experimental work, in the discussion of the next section the constant factors in Eq. (9) will be dropped, for convenience, and the relation taken to be:

$$G_3(\gamma)G_1(\gamma) = G_2(\gamma). \quad (9a)$$

Figure 1 illustrates the relations between the aperture field, the Fresnel field, and the distant fields.

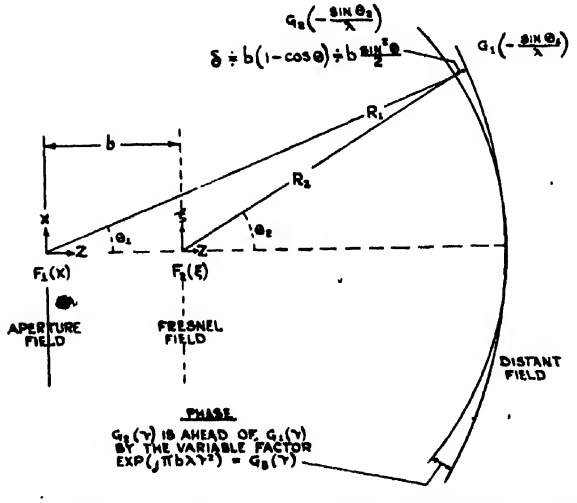


FIG. 1. Relations between the aperture field, Fresnel field and the distant field.

FOURIER TRANSFORM THEORY AND THE SOLUTION

For convenience in using tables[4] and to agree with Ramsay,[5] the definition has been made:

$$\gamma = -(\sin\theta)/\lambda$$

so that Eq. (2) has been written

$$G_1(\gamma) = \int_{-\infty}^{\infty} F_1(x) \exp(-j2\pi\gamma x)dx.$$

Michelson[6] in the field of optics of visible light and Spencer[7,8] and Ramsay[5] in connection with radiators have discussed the reciprocal relation of the Fourier transform pair that holds between an intensity distribution and the diffraction, or antenna pattern that it causes. These authors justify the use of the process inverse to that described in (2):

$$F_1(x) = \int_{-\infty}^{\infty} G_1(\gamma) \exp(j2\pi\gamma x)d\gamma. \tag{2a}$$

To abbreviate the analysis, transformation symbols are now introduced as has been done in the work of the writers,[5,7,8] who have been mentioned. Equations (2) and (2a) are written:

$$F_1(x) \leftrightarrow G_1(\gamma), \tag{10}$$

and in the same way:

$$F_2(\xi) \leftrightarrow G_2(\gamma), \tag{11}$$

$$F_3(\xi) \leftrightarrow G_3(\gamma) = \exp(j\pi b\lambda\gamma^2). \tag{12}$$

Since $G_2(\gamma) = G_1(\gamma)G_3(\gamma)$ it appears that on making use of the convolution theorem:

$$F_2(\xi) = \int_{-\infty}^{\infty} F_1(t)F_3(\xi - t)dt \leftrightarrow G_1(\gamma) \times G_3(\gamma). \tag{13}$$

$F_2(\xi)$ is the variable factor in the expression for the Fresnel field at $z = b$; Eq. (13) can be put into explicit form by first finding the transform of $F_3(\xi - t)$ from that of $F_3(\xi)$:

$$F_3(\xi - t) \leftrightarrow \exp(-j2\pi\gamma t)G_3(\gamma) = \exp(-j2\pi\gamma t) \exp(j\pi b\lambda\gamma^2) \tag{14}$$

and, by then writing out the integral implied by the transformation symbol:

$$F_3(\xi - t) = \int_{-\infty}^{\infty} \exp(j\pi b\lambda\beta^2) \times \exp(-j2\pi\beta t) \exp(j2\pi\beta\xi)d\beta \tag{15}$$

where the dummy label, β, has been introduced in place of, γ, to avoid confusion. Substitute (15) in (13):

$$F_2(\xi) = \int_{-\infty}^{\infty} F_1(t) \int_{-\infty}^{\infty} \exp(j\pi b\lambda\beta^2) \times \exp(-j2\pi\beta t) \exp(j2\pi\beta\xi)d\beta dt. \tag{16}$$

Equation (13) expresses the Fresnel field as a Fourier transform of the distant field and Eq. (16) expresses the Fresnel field in terms of the aperture field. Equation (9) and the reasoning that leads to it can be verified by evaluating the inner integral of Eq. (16) which then becomes:

$$F_2(\xi) = \left(\frac{j}{b\lambda}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} F_1(t) \exp\left(-j\frac{\pi}{b\lambda}[\xi - t]^2\right)dt. \tag{16a}$$

Equation (16a) will be recognized as the variable part of the Fresnel field expressed in the conventional manner by a Fresnel's integral and, in fact, (16a) can be made to agree completely with the conventional expression if the constant factors, which were dropped between Eqs. (9) and (9a), are reintroduced.

Up to this point there has been no suggestion of an obstruction in front of the aperture. Let a large screen be erected in the (ξ, η) plane at a distance, b, in front of S_1 and cut in it a rectangular opening, S_2, which extends along the ξ-axis from $-c/2$ to $+c/2$. Follow Kirchhoff in the assumption that in the opening S_2, $F_2(\xi)$ has its uninterrupted value, and that the intensity on the front surface of S_2 is zero everywhere else. $F_2(\xi)$ then must be given the mutilated form expressed by:

$$\begin{aligned} &F_2(\xi) = 0, && c/2 \leq |\xi| \leq \infty, \\ &F_2(\xi) \text{ is unchanged}, && -c/2 \leq \xi \leq c/2. \end{aligned} \tag{17}$$

The distant pattern will be modified from $G_2(\gamma)$ to a new function $G_{2m}(\gamma)$ by the mutilation of $F_2(\xi)$ which is caused by the screen. This can be written as the transform of $F_2(\xi)$ when expressed as in Eq. (16).

$$G_{2m}(\gamma) = \int_{-c/2}^{c/2} F_2(\xi) \exp(-j2\pi\xi\gamma)d\xi. \tag{18}$$

Reverse the order of integration so as to integrate first with respect to ξ, then with respect to t.

$G_{2m}(\gamma)$ is then found to be given by:

$$G_{2m}(\gamma) = c\int_{-\infty}^{\infty} G_1(\beta) \frac{\sin[\pi c(\beta - \gamma)]}{\pi c(\beta - \gamma)} \times \exp(j\pi b\lambda\beta^2)d\beta. \tag{19}$$

To convert $G_{2m}(\gamma)$ to $G_{1m}(\gamma)$, which is expressed as a wave diverging from S_1 (the original aperture), divide

[4] G. A. Campbell and R. M. Foster, "Fourier integrals for practical application," Bell Telephone System Monograph, B-584 (1931).
[5] J. F. Ramsay, Marconi Rev. IX, 139 (1946) ff.
[6] A. A. Michelson, Phil. Mag. 9, 506 (1906).
[7] R. C. Spencer, Radiation Lab. Report 762-1 (December, 1945).
[8] R. C. Spencer, Radiation Lab. Report 762-2 (March, 1946).

by $G_3(\gamma)$ as indicated by Eq. (9a):

$$G_{1m}(\gamma)=c\int_{-\infty}^{\infty}G_1(\beta)\frac{\sin[\pi c(\beta-\gamma)]}{\pi c(\beta-\gamma)}\times\exp(j\pi b\lambda[\beta^2-\gamma^2])d\beta. \quad (20)$$

Equation (20) when substituted in Eq. (1) gives the distant field of S_1 as modified by S_2 at $z=b$ when S_1 is not rotated and lies in the XY plane; the solution is in terms of the unmutilated pattern $G_1(\beta)$, the width, c, of S_2, the distance, b, between the two screens, and the wave-length, λ, of the radiation.

The effect of rotating S_1 out of the XY plane by an angle, α, measured between the normal to S_1 and the Z axis can be found as follows: If α is not too great, and b not too small, the effect of rotation is to introduce a linear phase shift into the original $F_1(x)$ distribution. This is expressed by:

$$F_{1\alpha}(x)=F_1(x)\exp[-j2\pi x(\sin\alpha)/\lambda] \quad (21)$$

which leads to the transformation:

$$F_1(x)\exp[-j2\pi x(\sin\alpha)/\lambda]\leftrightarrow G_1[\beta+(\sin\alpha)/\lambda]. \quad (22)$$

On substituting $G_1[\beta+(\sin\alpha)/\lambda]$ for $G_1(\beta)$ in Eq. (20) the complete solution of the problem is obtained:

$$G_{1\alpha m}(\gamma)=c\int_{-\infty}^{\infty}G_1\left[\beta+\frac{(\sin\alpha)}{\lambda}\right]\times\frac{\sin[\pi c(\beta-\gamma)]}{\pi c(\beta-\gamma)}\exp(j\pi b\lambda[\beta^2-\gamma^2])d\beta. \quad (23)$$

Equation (23) can be verified by reducing it to known forms. Put $\alpha=0$ and $b=0$ so that the second aperture S_2 lies on S_1 and limits $F_1(x)$ to the range $|x|\leq c/2$. Equation (23) becomes:

$$G_{1\alpha m}(\gamma)=c\int_{-\infty}^{\infty}G_1(\beta)\frac{\sin[\pi c(\beta-\gamma)]}{\pi c(\beta-\gamma)}d\beta, \quad (24)$$

which is a result obtained by Ramsay.[9]

If in Eq. (23) $c\to\infty$, then the screen S_2 no longer obstructs the near field. Since:

$$Lt_{c\to\infty}\frac{\sin[\pi c(\beta-\gamma)]}{\pi c(\beta-\gamma)}\frac{1}{c}=S_0(\beta-\gamma), \quad (25)$$

where $S_0(\beta-\gamma)$ is a unit impulse at $\beta=\gamma$, Eq. (23) reduces to the unmutilated distant field, $G_1[\beta+(\sin\alpha)/\lambda]$, when S_2 is removed from the near field.

[9] J. F. Ramsay, Marconi Rev. X, 90 (1947).

EXPERIMENTAL METHODS AND RESULTS

The amplitude of the electric field, in the aperture of an electromagnetic horn, in the E-plane, can be expressed to a good approximation by:

$$F_1(x)=F_A\exp[-j(\pi x^2/\lambda l)]$$

where

$$F_A=0 \text{ if } (a/2)<|x|, \qquad F_A=A \text{ if } |x|<a/2 \quad (26)$$

in which l = distance from the vertex to the aperture plane of the horn, a = width of the horn aperture in the E-plane. The E-plane aperture distribution function, for an electromagnetic horn, which may be rotated out of the XY plane as in Eq. (21) is therefore:

$$F_1(x)=F_A\exp\left(-j\left[\frac{\pi x^2}{\lambda l}+\frac{2\pi x\sin\alpha}{\lambda}\right]\right) \quad (27)$$

with F_A restricted as in (26).

Elliot and Rigby* have shown that on substitution of Eq. (27) in Eq. (18) it is possible to integrate Eq. (18) twice, then expand the resulting integral into an asymptotic series. They have estimated that when the constants of the expansion have the values $c=110$ cm, $b=100$ cm, $a=32$ cm, $\lambda=3.2$ cm, the first two terms of the series give the relative amplitudes of $G_{2\alpha m}(\gamma)$ to a precision of not worse than 1 db in the angular interval $-20°\leq\alpha\leq 20°$.

The E-plane amplitudes of $G_{2\alpha m}(\gamma)$ have been calculated at $\alpha=0$, 5, 10, 15, and 20 degrees for two horns, radiating 3.2-cm waves, through a mutilating aperture, 110 cm wide in the E-plane by 22 cm high in the H-plane, placed 100 cm in front of the horn aperture. The aperture of both horns was taken to be square, 32 cm to the side, and the lengths of the horns were taken as 100 and 50 cm, respectively. Patterns measured at a wave-length of 3.2 cm both with and without the mutilating aperture, for two horns of these dimensions are reported in Figs. 4 and 5 of a companion paper.[1] Comparison of the measured and calculated points, between $\pm 20°$ indicate that the probable error in the calculation, for the horn of length 100 cm is ± 0.25 db, and for the 50-cm horn, is ± 0.34 db.

ACKNOWLEDGMENT

The experimental work associated with this research was made possible by a grant for equipment from the Defence Research Board of the Department of National Defence, Ottawa, Canada. The writer wishes to thank Dr. H. A. Elliot and Miss E. C. Rigby, both of McGill University, for permission to refer to their unpublished expansion of Eq. (18).

* H. A. Elliot and E. C. Rigby (unpublished).

Second-Order Transition Temperatures and Related Properties of Polystyrene. I. Influence of Molecular Weight*

THOMAS G. FOX, JR., AND PAUL J. FLORY
Department of Chemistry, Cornell University, Ithaca, New York
(Received November 17, 1949)

Dilatometric and viscometric data on fractionated polystyrenes containing diethylbenzene end groups are presented over wide temperature ranges. The second-order transition temperature, viscosity-temperature coefficient, and specific volume all change rapidly with increasing molecular weight toward asymptotic limits which are practically reached at $M \simeq 30{,}000$. Empirical expressions are presented relating these properties to molecular weight and temperature. In each case the dependence on molecular weight is expressed as a simple function of $\bar{M}_n^{-1}$. These observations are interpreted and correlated on the basis of the hypothesis that the local configurational order in a liquid polymer is disturbed by the introduction of end groups to a degree that is proportional to their number. The second-order transition does not represent an isoviscous state. The internal local configurational structure appears to be equivalent, and independent of temperature, in all polystyrenes below their second-order transition temperatures.

INTRODUCTION

A LIQUID which may be cooled without incidence of crystallization solidifies to a glass over a fairly narrow temperature range which is characteristic of the substance.[1,2] This is true not only of hydrocarbons, polar and hydrogen-bonding liquids, inorganic glasses, etc., but also of amorphous polymers (including semi-crystalline, or semi-amorphous polymers) regardless of their composition or structural pattern (i.e., non-linear network polymers as well as linear ones). The temperature about which this transformation takes place is usually referred to as the "glass" temperature T_g, or the "second-order transition temperature"; for polymers it is often designated as the "brittle temperature," inasmuch as the polymer changes from a highly viscous liquid, or rubbery solid, to a brittle substance which will not tolerate appreciable deformation without fracture at all lower temperatures.

It is well established that this so-called "second-order transition" is not an equilibrium transition in the thermodynamic sense, but that it originates from kinetic limitations on the rates of the internal adjustments occasioned by changes in temperature (or by mechanical stress). At temperatures below T_g, the internal mobility is insufficient for maintenance of the equilibrium internal configuration and, hence, of the equilibrium volume. No discontinuity is observed in the first-order properties, volume and internal energy, but the second-order properties, thermal expansion coefficient and heat capacity, exhibit abrupt changes by means of which the "transition" is easily located.

The macroscopic viscosity often is employed as an inverse measure of internal mobility in monomeric substances, and it has been pointed out that the second-order transition generally occurs at a viscosity of about 10^{12} poises. The macroscopic viscosity of a polymer, however, depends not only on the mobility of each segment but also on an intersegment cooperation factor which increases rapidly with molecular weight.[3,4] If the second-order transition in polymers depends only on the local segment mobility and not on the interactions between all of the segments of a given molecule, then the magnitude of the macroscopic viscosity should not afford a universal measure of the approach to the second order transition; viscosities well above 10^{12} poises at T_g should be expected if the molecular weight is large. On the basis of the same postulate, the second-order transition might be expected to be independent of the molecular weight of the polymer except insofar as it affects the liquid configuration of the polymer.

Considerations such as these point to the desirability of securing reliable information on the dependence of the volume (as an index of liquid configuration) and of the viscosity of polymers on their molecular weights and temperatures, and on the dependence of T_g on molecular weight. Asymptotic increases of T_g's for polystyrenes,[5] polyisobutylenes,[6] and poly-(ethylacrylates)[7] with their "molecular weights" have been indicated by previous work on polymers of unknown degrees of heterogeneity. In a recent paper we have reported a qualitatively similar asymptotic increase in the viscosity-temperature coefficient with the molecular weights M of polystyrene fractions.[4] The rapid increase of both T_g and the viscosity-temperature coefficient toward their limiting values with increase in M contrasts sharply with the sustained increase in viscosity with increasing M even to the highest values investigated.[4]

* The work presented in this paper was carried out at Cornell University in connection with the Government Research Program on Synthetic Rubber under contract with the Office of Rubber Reserve, Reconstruction Finance Corporation.

[1] W. Kauzmann, Chem. Rev. 43, 219 (1948).
[2] R. F. Boyer and R. S. Spencer, *Advances in Colloid Science* (Interscience Publishers, Inc., New York, 1946), Vol. II, p. 1.
[3] P. J. Flory, J. Am. Chem. Soc. 62, 1057 (1940).
[4] T. G. Fox, Jr. and P. J. Flory, J. Am. Chem. Soc. 70, 2384 (1948).
[5] E. Jenckel and K. Ueberreiter, Zeits. f. physik. Chemie A182, 361 (1938).
[6] K. Ueberreiter, Zeits. f. physik. Chemie B45, 25 (1940).
[7] R. H. Wiley and G. M. Brauer, J. Polymer Sci. 3, 647 (1948).

TABLE I. The polystyrenes used for fractionation.

Polymer	Molar ratio styrene/diethylbenzene	Reaction temperature	Reaction time, hours	Extent of reaction (percent)	$\bar{M}_v$	Initial conc. for fractionation g/100 ml
A	1	150°	2	27	120,000	2.0
B	0.093	150°	9	72	9560	2.0
C	0.051	150°	17	90	4870	1.5

In the present investigation the specific volumes of polystyrene fractions have been investigated as functions of temperature and molecular weight. Second-order transition temperatures have been secured in the customary fashion from the discontinuity in the thermal expansion coefficient. The results have been correlated with the viscosity-temperature coefficient and its observed dependence on M. Polystyrene has been chosen as a model non-polar polymer for this investigation on account of the convenient location of its second-order transition temperature, the simplicity of its structure, and the comparative abundance of information concerning its various properties.

EXPERIMENTAL

Preparation and Fractionation

Three polymers of styrene of varying molecular weights (Table I) were prepared by thermal polymerization in diethylbenzene *in the absence of added catalyst.* The concentrations of styrene employed and other details pertinent to these polymerizations are summarized in Table I. The polymers were separated by precipitation in an excess of vigorously stirred methanol and were dried in vacuum at 60°.

These polymers were fractionated by the single precipitation technique described in a previous paper.[4] The initial concentrations of the solutions employed are shown in the last column of Table I. Each of the two polymers of lower molecular weight was separated into nine fractions by fractionating from acetone solution at 30° using as the precipitant a mixture consisting of equal volumes of water and methanol. Each individual fraction employed here represented from 5 to 17 percent of the whole polymer (Table II). The polystyrene of higher molecular weight (120,000) was separated into three fractions employing methyl ethyl ketone-methyl alcohol as the solvent-precipitant combination. Only the middle fraction, $M=85{,}000$, representing 38 percent of the whole polymer was retained for subsequent experiments.

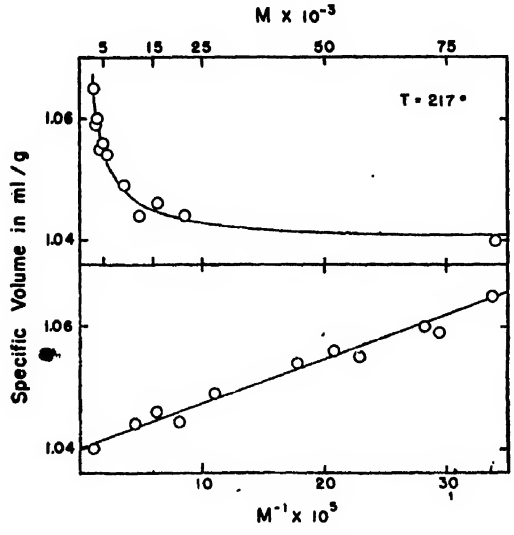


FIG. 1. Specific volume at 217° for polystyrene fractions *vs.* molecular weight (upper curve) and *vs.* M^{-1} (lower curve).

Molecular Weight Determination

Molecular weights of the polystyrene fractions have been calculated from their intrinsic viscosities, $[\eta]$, in freshly distilled benzene, using the relationship of Ewart[8]

$$\log\bar{M}_v=(\log[\eta]+4.013)/0.74,$$

where $\bar{M}_v$ is the viscosity average molecular weight. The solution viscosities were measured at 30.00±0.01° using a Ubbelohde No. 1 viscometer the calibration of which included the kinetic energy term.[9] In each case the intrinsic viscosity has been calculated from the relative viscosity, η_r, of a single solution by means of the viscosity-concentration relationship reported by Tingey.[10]

$$\ln\eta_r/c=[\eta]-0.125[\eta]^2c,$$

where c is expressed in g/100 ml.

Molecular weights determined in this way are generally reproducible to ±2 percent.

Determination of Specific Volumes

The pycnometers and the technique employed were similar to those described in a previous publication.[3] Densities were determined at 217° by filling a weighed pipette of known volume (from 1 to 2 ml) with the liquid polymer at this temperature, and subsequently weighing at room temperature. The reproducibility of the data was generally within ±0.1 percent.

Dilatometric Measurements

The dilatometers and techniques employed resemble those described by Bekkedahl.[11] A weighed quantity (generally about two grams) of bubble-free polymer, previously dried under vacuum at 217°, was placed in the bulb of a glass dilatometer and the bulb was sealed, with precautions to avoid heating the polymer. Mercury was introduced under vacuum (0.005 mm Hg).

[8] R. H. Ewart, paper presented at the Atlantic City Meeting of the American Chemical Society (April 14, 1947).

[9] The calibration method outlined in ASTM Designation D445–39T was employed.

[10] H. C. Tingey, unpublished data referred to by R. H. Ewart, *Advances in Colloid Science* (Interscience Publishers, Inc., New York, 1946), Vol. II, p. 210.

[11] N. Bekkedahl, J. Research Nat. Bur. Stand. 43, 145 (1949).

TABLE II. Dilatometric data on polystyrene fractions.

Polymer fraction	Percent of unfractionated polymer	M	T_g	v_{140} ml/g	v_g ml/g Obs.	v_g ml/g Calc. by Eq. (4)	$(dv/dT)_1$	$(dv/dT)_2$, ml/g/° ×10⁴	$(dv/dT)_3$
*A*2	38	85000	100°	0.989	0.969	0.968	–	4.9	2.3
*A*2	38	85000	100	0.990	0.969	0.968	7.2	5.1	2.0
*A*2	38	85000	98	0.989	0.970	0.968	—	4.6	2.2
*A*2	38	85000	99	0.991	0.970	0.968	7.1	5.0	2.4
*B*1	10.0	22000	89	0.992	0.963	0.965	—	5.7	2.7
*B*2	15.0	16000	86	0.993	0.962	0.964	—	5.7	2.3
*B*4	16.7	9150	77	0.996	0.960	0.962	—	5.6	2.4
*B*5	12.5	7200	78	0.998	0.964	0.962	7.0	5.5	—
*B*6	10.0	5600	73	1.003	0.963	0.960	—	6.1	2.6
*B*6	10.0	5600	75	1.000	0.967	0.960	—	5.1	2.7
*B*6	10.0	5600	76	1.002	0.966	0.960	7.3	5.6	—
*C*6	16.7	4810	64	1.003	0.962	0.959	—	5.3	2.6
*C*6	16.7	4810	66	1.003	0.962	0.959	7.6	5.5	-
*B*7	5.4	4300	62	1.003	0.956	0.958	—	6.0	2.4
*B*7	5.4	4300	63	1.003	0.957	0.958	—	5.8	2.5
*B*7	5.4	4300	62	1.002	0.956	0.958	7.6	6.0	—
*B*7	5.4	4300	63	1.003	0.957	0.958	—	5.9	2.8
*C*7	12.3	3540	53	1.005	0.957	0.956	—	5.5	2.6
*C*8	10.5	2970	40	1.007	0.955	0.953	–	5.3	2.1
*C*8	10.5	2970	—	1.005	0.953	0.953	8.4	5.3	—
*C*8	10.5	2970	—	1.006	0.953	0.953	8.0	5.3	—
*C*8	10.5	2970	-	1.007	0.952	0.953	—	5.5	—

The amount of mercury in the dilatometer was determined by weighing. The volume of polymer plus mercury was calculated at the calibration temperature, 217°, where the density of the polymer was known from the absolute specific volume measurements. The position of the mercury in the capillary tube was noted at 217° and then at lower temperatures usually at intervals of 10°. The measurements were made in a constant temperature bath regulated within ±0.1° and the relative densities generally were accurate to ±0.03 percent. At all temperatures except those in the vicinity of T_g, volume equilibrium was quickly established and in trial experiments no further change could be observed over a period of 24 hr. Hence before recording the volume the sample was held for a period of 10 to 60 min. at each temperature 10° or more above T_g and from 6 to 12 hr. at all lower temperatures.

Melt Viscosities

Melt viscosities in the range of 10 to 10^6 poises were determined by observing the rate of flow into uniform bore capillary viscometers under a measured pressure differential. The calibration and use of these viscometers have been described.[4] For measurement of viscosities from 0.1 to 10 poises, pipette type viscometers were employed. These consisted in each case of a Pyrex capillary tube of 1-mm bore with a small bulb (capacity 0.2 to 1.0 ml) about 12 to 15 cm from the lower end of the capillary. Absolute viscosities in poises were calculated in this case from $\eta = klp$, where t is the measured time required to fill the viscometer bulb under a predetermined pressure differential p, and k is a constant determined from the tp product for an oil the viscosity of which was accurately determined with the capillary viscometers mentioned above. Temperature control was provided by means of vapor baths (±0.3°) above 100° and by thermostated oil baths (±0.1°) at lower temperatures. The viscosity values usually were reproducible to ±3 percent.

High Temperature Stability

It was reported previously[4] that phenyl β-naphthylamine greatly retards the degradation of polystyrene at high temperatures, and that 0.5 percent of this substance assures adequate stability to permit measurement of viscosities up to 217°. The experiments on which this conclusion was based were conducted on polymers of high molecular weight, the assumption being made that the lower polymers would be at least

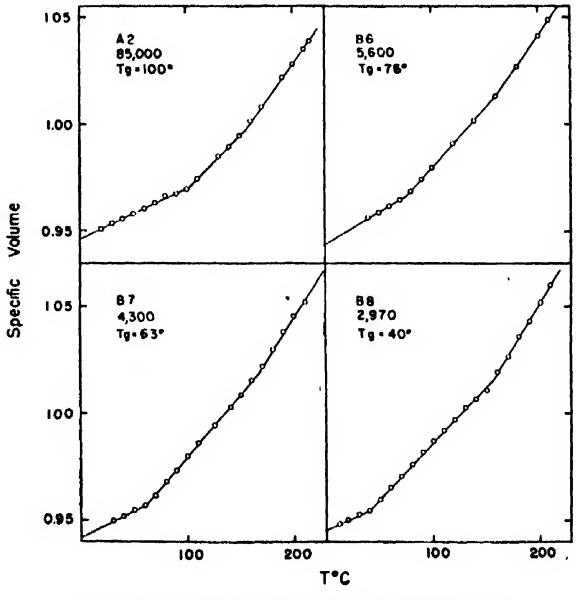


FIG. 2. Specific volume-temperature curves for polystyrene fractions.

TABLE III. Viscosity-temperature data for polystyrene fractions.

Polymer fraction	M	η_{217}, poises	log(η_T/η_{217}) at the indicated temperature (°C) 190	160	140	130	120	110	100	95	90	85	80	75
*A*2	85000	2000		2.0										
*B*1	21500	51.3	0.70	1.94	3.36	4.10								
*B*2	16000	26.7	0.74	1.85	3.21	3.89								
*B*3	12300	15.7												
*B*4	9100	9.13	0.54	1.61	2.80[a]	3.45	4.31							
*B*5	7230	5.29												
*B*6	5600	2.60	0 50	1.44		2.99		4.58	5.72					
*C*6	4810	1.36	0.45	1.25	2.17			4.03	4.95	5.52				
*B*7	4300	1.55	0.38	1.26		2.53		4.06	4.96					
*C*7	3540	0.59	0.47	1.11	1.89			3.54	4.29		5.24	5.83		
*B*8	3400	0.80	0.42	1.01		2.16			4.19		5.25			
*C*8	2970	0.32	0.41	0.93	1.58			3.07	3.58		4.38		5.33	5.92

[a] Interpolated.

as stable. Subsequent experiments revealed, however, that phenyl β-naphthylamine fails to inhibit degradative reactions in the lower polystyrenes included in the previous investigation. These low polymers were prepared with the aid of benzoyl peroxide initiation, and a connection between the observed instability and the presence of benzoate end groups in the polystyrene molecules[12] was indicated by the sublimation of a quantity of pure benzoic acid when these polymers were heated at 217° under reduced pressure. The low molecular weight polystyrenes prepared as described above *without the use of catalyst* (Table I) were relatively stable. When heated for an hour at 217° in an inert atmosphere (pure nitrogen) they discolored only slightly and exhibited no change in intrinsic viscosity even in the absence of inhibitor.

RESULTS

Specific Volumes at 217° *vs.* Molecular Weight for Polystyrene Fractions

The specific volumes of eleven polystyrene fractions ranging in molecular weight from 2970 to 85,000 were determined at 217° by the constant volume pipette method. The specific volume at 217°, v_{217}, decreases with increasing molecular weight in the lower range and is relatively constant for molecular weights above 25,000 (upper curve of Fig. 1). When the specific volume is plotted against $1/M$ as in the lower portion of Fig. 1, the data are well represented by the straight line, the equation of which is

$$v_{217}=1.040+72/M. \qquad (1)$$

The units employed for the specific volume throughout this paper are ml/g.

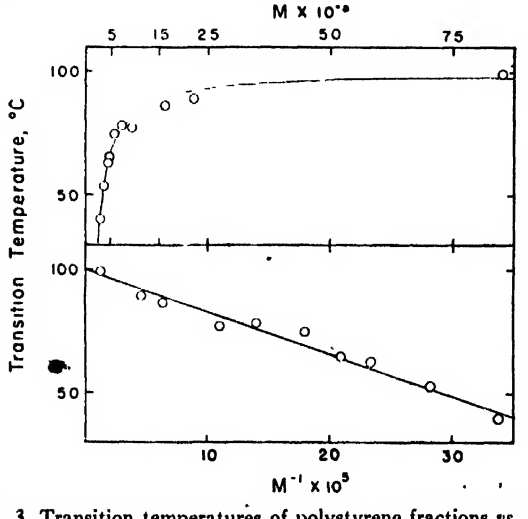


FIG. 3. Transition temperatures of polystyrene fractions *vs.* molecular weight (upper curve) and *vs.* M^{-1} (lower curve).

[12] P. D. Bartlett and S. G. Cohen, J. Am. Chem. Soc. 65, 543 (1943).

Dilatometric Data on Polystyrene Fractions

Specific volumes were measured by the dilatometric method as a function of temperature on ten polystyrene fractions ranging from 2970 to 85,000 in molecular weight. As illustrated in Fig. 2, the volume-temperature curve up to 160° for each polymer may be represented by two intersecting straight lines. The temperature at which these lines intersect is taken as the second-order transition temperature, T_g. In the vicinity of 160°, the volume-temperature coefficient (i.e., the slope) exhibits a further increase with increasing temperature. The results in this range may again be approximated, for simplicity, by two intersecting straight lines, the temperature at the intersection being about 160°. Hence, each polymer may be characterized by five quantities: T_g, the volume v_g at T_g, and three volume-temperature coefficients, $(dv/dT)_1$ above 160°, $(dv/dT)_2$ between 160° and T_g, and $(dv/dT)_3$ below T_g. The values of these parameters, as obtained from a series of experiments, details for which are not included here, are presented in Table II.

The transition temperature T_g becomes almost independent of molecular weight above 25,000 but decreases rapidly with decreasing molecular weight below this value (upper curve of Fig. 3). In analogy with the dependence of specific volume on M^{-1}, T_g is likewise a linear function of M^{-1}, as is shown in the lower curve of Fig. 3. The equation of the straight line is

$$T_g=100-1.7\times10^5/M \qquad (2)$$

from which the upper curve of Fig. 3 has been calculated.

Although at T_g the abrupt change in slope and the accompanying increase in the time needed for equilibration justify the use of a function with a discontinuity in its slope at this temperature, the use of two straight lines intersecting at 160° to represent the data above T_g is arbitrary. The change of slope at 160° is much smaller than that observed at T_g, and a line with curvature over a wide range about 160° would represent the data equally well.

The specific volume-temperature coefficients $(dv/dT)_2$ and $(dv/dT)_3$ (Table II) appear to be independent of molecular weight within the admittedly large experimental error involved in their evaluation over limited temperature ranges. Average values are 5.5 and 2.5 $\times 10^{-4}$, respectively. The value of $(dv/dT)_1$ is about 7×10^{-4} with an indication of an increase at very low molecular weight.

The portion of the volume-temperature curve between T_g and 160° may be represented by the equation

$$v = A + 5.5\times 10^{-4}T + 53/M, \quad 160 \geq T \geq T_g, \qquad (3)$$

where A is a constant equal to 0.913 and T is in degrees centigrade. Combining this with Eq. (2), we obtain

$$v_g = 0.943 + 2.4\times 10^{-4}T_g. \qquad (4)$$

Values of v_g calculated by this equation agree within experimental error with the observed values, as shown in Table II. In view of the non-dependence of $(dv/dT)_3$ on molecular weight, the specific volume below T_g is given by

$$v = v_g - 2.5\times 10^{-4}(T_g - T), \quad T \leq T_g$$

and combining this with Eq. (4), ignoring the slight difference between the two temperature coefficients, we obtain

$$v = 0.943 + 2.5\times 10^{-4}T, \quad T \leq T_g. \qquad (5)$$

This equation illustrates the important observation that the specific volume below the transition temperature is substantially independent of the molecular weight.

Volume-Time Relationships

The effect of temperature on the rate of volume relaxation was determined for several polystyrene fractions exhibiting values of T_g from 60 to 100°. In each case a dilatometer containing the polymer (and mercury) was quickly transferred from a bath at temperature T_1, generally about 30° above T_g, to a second constant temperature bath at a lower temperature T_2. The volume was then observed as a function of time. Subsequently the polymer was again heated to T_1 and the procedure repeated at a different temperature T_2. At temperatures T_2 exceeding T_g by about 10°, temperature and volume equilibria appeared to be attained simultaneously. At lower temperatures the volume was observed to decrease approximately linearly with the logarithm of time, at least for periods up to 24 hr. The magnitude of the rate of volume relaxation, which was too high to measure about 10° above T_g, had a value of about 30 to 50×10^{-4} at T_g, as measured by the slope of the v *vs.* log$t$ curve (t in min.), and decreased rapidly with decreasing temperature, becoming practically constant at a temperature generally about 10° below T_g. This limiting value of $-dv/d\log t$ was about 8×10^{-4} for all polystyrenes studied.

These results emphasize the fact that although the value of T_g depends on the observation time employed, the marked temperature dependence of the rate of volume relaxation in the vicinity of T_g makes it impossible, using reasonable periods of time, to appreciably alter the value of T_g; it would, for example, be increased by only 5 or 10° on decreasing the time scale of the experiment from 24 hr. to 5 min.

It has been reported[13] that high molecular weight polystyrene previously kept at 20° for two years can be heated from 20 to 140° without undergoing a second-order transition and, further, that the annealing times for this polymer at 20 to 30° below the usual value of T_g are measured in hours or at most in a few days. This reported observation cannot be reconciled with the present results which indicate that the annealing time in this range would consist of many centuries at the least. It is therefore doubtful whether an equilibrium volume-temperature curve for polystyrene has been obtained in the indicated range.[14]

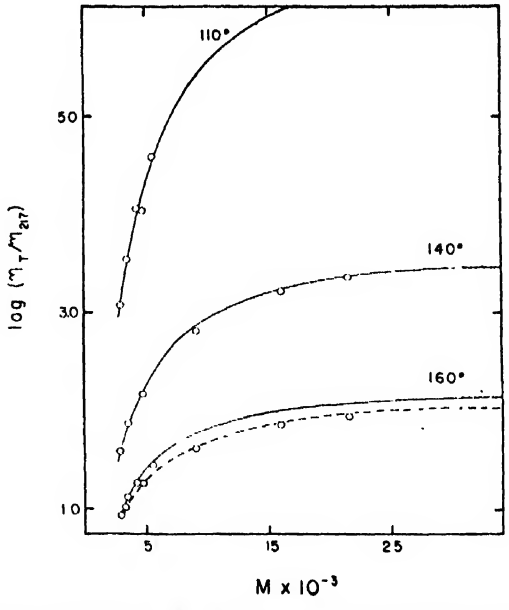


FIG. 4. Log(η_T/η_{217}) *vs.* M for polystyrene fractions at several temperatures. The solid curves are calculated by Eq. (7). The dotted line is calculated by Eq. (6) using the value of $F(T)$ at 160° obtained from Fig. 5.

[13] R. S. Spencer and R. F. Boyer, J. App. Phys. 17, 398 (1946).
[14] Close scrutiny indicates that the reproducibility of the data of reference 13 is poor. In at least one of the two experiments the

Viscosity-Temperature-Molecular Weight Relationships for Polystyrene Fractions

Viscosities were measured at 217° and at lower temperatures for twelve polystyrene fractions of molecular weight from 2970 to 85,000 (Table III). The viscosities at 217° (η_{217}) for the fractions of higher molecular weight are in good agreement with the viscosities previously reported,[4] while those for the fractions of lower molecular weight are generally about 10 percent higher. This discrepancy almost certainly arises from the susceptibility to decomposition of the lower polystyrenes previously used as pointed out in the Experimental section. The relation between $\log(\eta_T/\eta_{217})$ and $1/T$ (in °K) is non-linear and its dependence on molecular weight is qualitatively similar to that previously reported.[4,15]

The dependence of the viscosity-temperature coefficient on molecular weight is illustrated in Fig. 4 where values of $\log(\eta_T/\eta_{217})$ at 160, 140, and 110° are plotted *vs.* the molecular weight. At a given temperature T the quantity $\log(\eta_T/\eta_{217})$ increases with increasing molecular weight up to about 25,000 beyond which it is relatively constant. Its dependence on M is similar for different temperatures, which suggests that $\log(\eta_T/\eta_{217})$ may be expressed as the product of two factors, one of which depends only on T and the other only on M. In further support of this supposition, and in analogy with the dependence of the specific volume and the second-order transition temperature on M, plots (Fig. 5) of $\log\log(\eta_T/\eta_{217})$ *vs.* $1/M$ from 90 to 190° are represented by a set of parallel straight lines. The indicated relationship may be written

$$\log(\eta_T/\eta_{217}) = F(T)e^{-2530/M} \qquad (6)$$

where $F(T)$ is a function of the absolute temperature only. This equation offers a satisfactory limiting form at both very low and very high molecular weights.[16]

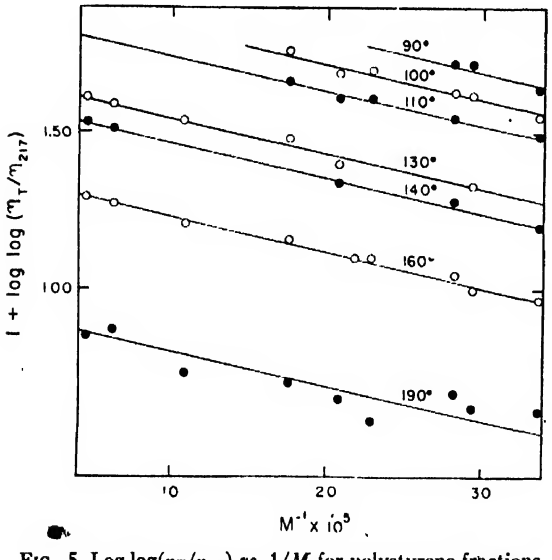


FIG. 5. Log $\log(\eta_T/\eta_{217})$ *vs.* $1/M$ for polystyrene fractions at various temperatures.

In the quest for a suitable empirical expression for $F(T)$, the viscosity-temperature function at $M=\infty$, it has been found (Fig. 6) that the values of $F(T)$ obtained from the intercepts at $1/M=0$ in Fig. 5 are linear functions of $1/T^6$, where T is in °K.[17] Hence the empirical relationship

$$\log(\eta_T/\eta_{217}) = 2.92\times10^{15}(1/T^6 - 1/\langle 490\rangle^6)e^{-2530/M} \qquad (7)$$

expresses the dependence of the temperature coefficient of viscosity on both temperature and molecular weight. Differentiating we obtain the expression for E_T, a quantity which we have called the "apparent energy of activation for viscous flow."[4]

$$Rd\ln(\eta_T/\eta_{217})/d(1/T) = E_T = (8.1\times10^{17}/T^5)e^{-2530/M} \qquad (8)$$

where the units for E_T are cal./mole.

The extent to which the data of Table III conform to Eq. (7) is illustrated in Fig. 4. The solid curves are calculated by Eq. (7) and the circles represent the experimental points. At 110 and 140° the agreement is good, while at 160° the calculated ratio (solid line) is too high. This failure is due to the inadequacy of the form of the temperature dependence rather than to the exponential dependence on the molecular weight, as is indicated by the agreement between the dotted line of Fig. 6 and the experimental points at 160°. The latter line was calculated using the value of $F(T)$ obtained from the extrapolation of the data at 160° to $M=\infty$ (Fig. 5) rather than from Eq. (7). [The values of $F(T)$ at 140 and 100° obtained from the intercepts of Fig. 5 are in agreement with the respective values calculated for $M=\infty$ by Eq. (7).]

In general the discrepancy between viscosities calculated by Eq. (7) and the observed values is less than 20 percent. It is possible that the higher discrepancies which occur are due to a slight degradation of the polymer during the viscosity determinations. On the other hand, Eqs. (7) and (8) are presented merely as approximate empirical representations of the data, which may not necessarily be significant theoretically. They are included primarily for their value in interpolating and extrapolating the experimental measurements.

data could be represented as having a second-order transition at 90°. It may be significant that the values of the "equilibrium" expansion coefficients were about 25 to 30 percent lower than that corresponding to the value of $(dv/dT)_2$ reported here.

[15] The viscosity-temperature coefficients for the high molecular weight polystyrene fractions are identical with those reported in reference 4, but they decrease less rapidly with decreasing molecular weight. These differences probably are due to the occurrence of decomposition in the previous unstable fractions.

[16] Over the molecular weight range investigated, $\log(\eta_T/\eta_{217})$ may alternately be represented as decreasing linearly with $1/M$. However, this relationship must fail at low molecular weights when the quantity $(1-\text{const.}/M)$ becomes small or negative (const.$\cong$1600).

[17] The use of the sixth power of the temperature affords a better approximation to the data than either the fifth or the seventh powers. No attempt was made to employ a fractional power of T.

The use of these equations is illustrated in Table IV where values of $\log\eta$ and E_T have been calculated for polystyrenes of various molecular weights at their transition temperatures. Here T_g has been calculated in each case from the molecular weight by Eq. (2); $\log\eta_T$ at T_g has been calculated from Eq. (7) together with the viscosity-molecular weight relationship at 217°,[4] and E_T has been calculated by Eq. (8). The viscosities of polystyrene fractions at their transition temperatures are not identical but on the contrary vary considerably with molecular weight. The calculated values of E_T at T_g for the various polystyrene fractions, however, are all within 2 kcal. of 110 kcal. per mole. The second-order transition temperatures for a series of polystyrenes do *not* occur at an isoviscous state but *the viscosity-temperature coefficient at the transition temperature is independent of molecular weight.*

The approximate constancy of E_T at T_g may be deduced alternately from Eq. (8). If we require the temperature T at which a polystyrene of molecular weight M has the same value of E_T as that for polystyrene of $M=\infty$ at the transition temperature of the latter ($T_{g,\infty}$), then from Eq. (8)

$$\frac{8.1\times10^{17}}{T^5_{g,\infty}}=\frac{8.1\times10^{17}}{T^5}\,e^{-2530/M}$$

or solving for T and using the first two terms of the series expansion of the exponential term

$$T=T_{g,\infty}(1-506/M). \qquad (9)$$

Inserting the value of 373°K for $T_{g,\infty}$, and converting to °C

$$T=100-1.9\times10^5/M \qquad (10)$$

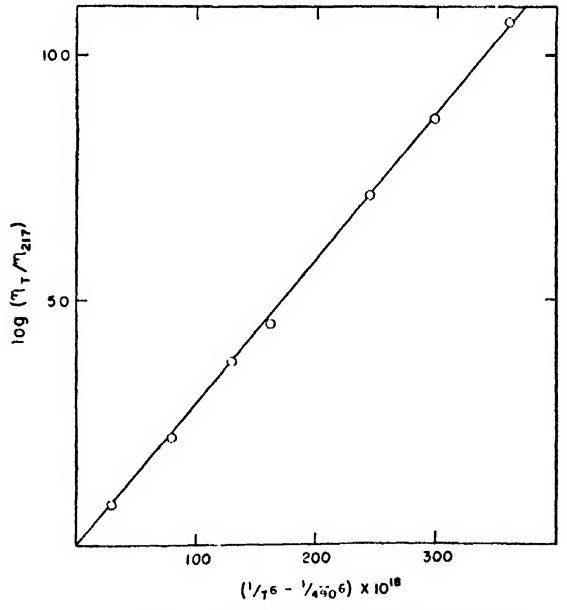


FIG. 6. Log (η_T/η_{217}) for polystyrene of infinite molecular weight *vs.* ($1/T^6-1/(490)^6$).

TABLE IV. Viscosities and viscosity-temperature coefficients of polystyrene fractions at their transition temperatures.

M	T_g	$\log\eta$ at T_g	E_T at T_g kcal./mole
300,000	99°	14.19	112
100,000	98	12.40	112
50,000	97	11.40	111
25,000	93	10.93	111
15,000	89	10.57	110
10,000	83	10.58	110
5000	66	10.66	109
4000	58	10.69	108
3000	43	11.31	111

which is identical with Eq. (2) within the limits of experimental accuracy and the approximations involved.

Mixtures of Polystyrene Fractions

Equations (1), (3), (2), and (7) relating the specific volume, the transition temperature, and the viscosity-temperature coefficient to the molecular weight M have been established from measurements on polystyrene fractions which, for the purpose of the experiments described, may be considered to be essentially homogeneous in molecular weight. In order to establish the particular average (if any) on which these properties depend, three mixtures of polystyrene fractions were prepared having the compositions given in Table V. The values of $\bar{M}_n$ and $\bar{M}_w$ calculated by the usual equations

$$\bar{M}_n=\sum_i w_i/\sum_i(w_i/M_i) \qquad (11)$$

$$\bar{M}_w=\sum_i w_i M_i/\sum_i w_i \qquad (12)$$

where w_i is the weight fraction of the species of molecular weight M_i are also listed in Table V. In each case a benzene solution containing the indicated proportions of the two fractions of widely separated molecular weights was evaporated to dryness in order to assure intimate mixing of the constituents. Measured specific volumes, transition temperatures, and viscosity-temperature coefficients are compared in Table V with the values computed from Eqs. (1), (2) and (7), using the number and the weight average molecular weights of each mixture. In every case the observed value of the given property agrees more nearly with that calculated for a fraction possessing a molecular weight which matches the number rather than the weight average molecular weight of the mixture. Agreement with the calculations based on the number average molecular weight is in all cases within the experimental error. Hence, for a heterogeneous polystyrene the specific volume, transition temperature, and the viscosity-temperature coefficient[18] depend explicitly on the number average molecular weight of the polymer, in accordance with Eqs. (1), (2), (3), (7), and (8).

The volume-temperature coefficients observed for the

[18] The dependence of the viscosity-temperature coefficient for a heterogeneous polystyrene on $\bar{M}_n$ was reported in reference 4.

TABLE V. Mixtures of polystyrene fractions.

A. Molecular weights and specific volumes

Designation	Components	$\bar{M}_w$, calc.	$\bar{M}_n$, calc.	v_{217} ml per g: For fraction of same $\bar{M}_w$	v_{217} ml per g: For fraction of same $\bar{M}_n$	Obs.
1	85,000(0.5) 4810(0.5)	45,000	9110	1.042	1.048	1.046
2	85,000(0.5) 3540(0.5)	44,300	6800	1.042	1.051	1.048
3	85,000(0.25) 3540(0.75)	23,900	4650	1.043	1.055	--

B. Second-order transition temperature and volume temperature coefficients

Designation	T_g: For fraction of same $\bar{M}_w$	T_g: For fraction of same $\bar{M}_n$	Obs.	(dv/dT) in ml/g/°×10⁴: $(dv/dT)_1$	$(dv/dT)_2$	$(dv/dT)_3$
1	96	81	80	6.8	5.3	2.4
2	96	75	70	6.8	5.5	2.3
3	93	63	60	6.8	5.5	2.6

C. Viscosity-temperature coefficients

Designation	η_{217}, poises	$\log(\eta_{160}/\eta_{217})$: For fraction of same $\bar{M}_w$	For fraction of same $\bar{M}_n$	Obs.	$\log(\eta_{140}/\eta_{217})$: For fraction of same M_w	For fraction of same M_n	Obs.
1	128	2.19	1.76	1.64	3.57	2.85	2.83
2	88	2.19	1.60	1.54	3.57	2.60	2.63

mixtures (Table V) are similar to those found for the fractions.

DISCUSSION

According to current concepts,[19] the equilibrium configuration of a liquid changes continuously with temperature in a manner such that the density of packing of the molecules, or of the chain segments in the case of a polymer, increases with decreasing temperature. The degree of order also increases as the temperature is lowered, and the character of the spatial arrangement, as expressed for spherical molecules by the average coordination number, may change as well.

The linear increases in the specific volume v of polystyrene with $1/\bar{M}_n$ may be interpreted as an influence of the end groups which is proportional to their concentration. Since the end groups, i.e., the $C_2H_5—C_6H_4—\overset{CH_3}{\underset{|}{CH}}—$ groups, occurring in the polystyrenes used here and the styrene structural unit should possess similar net densities, we are led to suggest that the observed dependence of specific volume on molecular weight reflects changes in average density of packing with the concentration of end groups. Thus, the end groups act like a foreign substance in disrupting the local configurational order of the styrene units, and this is manifested by an increase in the specific volume v which is proportional to the concentration of end groups, or to $1/\bar{M}_n$. It follows also that the specific volume affords a convenient index of the average internal configuration or of the average density of packing irrespective of M.

The specific volume-temperature relationships for polystyrene are summarized in Fig. 7. Three parallel straight lines represent the equilibrium v-T relationships for polystyrenes of different molecular weights. The broken line intersecting these v-T lines represents the barrier imposed by the second-order transition. The dotted extensions of the v-T lines into the region below the second-order transition are essentially unrealizable experimentally. All polystyrenes whose *equilibrium* v-T lines would place them in this forbidden region exhibit the *same specific volume-temperature relationship*, as given by the broken line, *independent of molecular weight*. The position of the broken line representing the glassy state depends somewhat on the rate at which observations are made, but, as has been pointed out above, this rate must be altered by several orders of magnitude before it is shifted appreciably.

The internal changes in liquid configuration which must occur on lowering of the temperature obviously involve the motion of segments relative to their nearest neighbors from one configurational arrangement to another. When the temperature is altered well above T_g this motion of the segments relative to their neighbors in the relatively "loose" or "open" structure is rapid and volume equilibration is practically instantaneous. As the temperature approaches T_g however, the configurational structure becomes more and more compact and hence the rate of volume equilibration diminishes rapidly. Finally, at T_g the segments are so densely packed that the internal mobility is negligibly small and consequently, the liquid structure is "frozen-in," and therefore *remains the same for all temperatures below* T_g. The identity of the specific volumes for polystyrenes of different molecular weight at all temperatures below this second-order transition leads at once to the conclusion that *below* T_g *the local configurational arrangement of the polymer segments is independent of both* M *and* T. Furthermore, it is clear that since the presence of end groups in the liquid polymer at a given temperature decreases the density of the configurational structure in proportion to $1/\bar{M}_n$ and since lowering the temperature increases this density of packing in proportion to T, the temperature at which this critical

[19] J. D. Bernal, Trans. Faraday Soc. 33, 27 (1937).

structure of low mobility is reached (T_g) must decrease in a linear fashion with $1/\bar{M}_n$.

Although the configurational structure remains unchanged as T is lowered below T_g, there remains a small linear decrease in volume with temperature associated with the crystal-like "contraction of the lattice" due to decreased amplitudes of oscillation of the segments about mean fixed positions. Obviously the equilibration rate for this process should be very rapid and should be essentially independent of the structural details mentioned above.

If a model for the liquid is chosen in which the chain segments are represented as hard, inelastic bodies, in accordance with certain approximate theories of the liquid state,[20,21] the above conclusions translate to the statement that the "free volume" (or volume of "holes" between the segments) in polystyrene below its transition temperature is a constant independent of both molecular weight and temperature. This follows from the separation of the specific volume into the sum of the free volume v_f and the occupied volume v_0, the latter being given by

$$v_0 = A_0 + (dv/dT)_3 T, \tag{13}$$

where A_0 is the hypothetical volume of the solid (glass) at $T=0$°C. Subtracting this expression from Eq. (3) for the temperature range from T_g to 160° we obtain

$$v_f = A_f + [(dv/dT)_2 - (dv/dT)_3]T + 53/M, \tag{14}$$

where $A_f = A - A_0$. Imposing the requirement that the equilibrium free volume of a polystyrene of molecular weight M at its transition temperature T_g be equal to that of one for which $M=\infty$ at its transition temperature $T_{g,\infty}$, then

$$T_g = T_{g,\infty} - 53/[(dv/dT)_2 - (dv/dT)_3]M. \tag{15}$$

Substituting the values of the parameters given above,

$$T_g = 100 - 1.8\times 10^5/M, \tag{16}$$

which is identical within the limits of experimental accuracy with Eq. (2). Hence, the second-order transition for polystyrene represents a state of "*iso-free volume*," and this same free volume is retained at all temperatures below T_g.

The above considerations suggest that the internal relaxation rate (mobility) is determined almost entirely by the configurational arrangement of the segments (or by the free volume associated with the given configuration) and that it is otherwise very nearly independent of both temperature and molecular weight. A corresponding dominant role of the volume in determining the internal mobility of simple non-polar liquids is illustrated by the fact that their fluidities depend primarily on the specific volume, i.e., the viscosity-temperature coefficient at constant volume is nearly zero.[22] As the temperature is lowered (for any liquid substance) toward T_g, the free volume shrinks to a small value and below T_g remains constant for all lower temperatures. Inasmuch as the volume decreases about linearly with temperature, the *percentage* change in free volume per degree increases rapidly as the temperature is lowered near T_g. Both the observed precipitous decrease of $-dv/d\log t$ for polystyrene near T_g and the approximate constancy of this coefficient below T_g follow as consequences of these considerations.

These concepts may be re-stated in mathematical form. The observations cited[22] on the fluidities of simple liquids suggest that the viscosity of such a substance may be expressed as the product of two functions as follows:

$$\eta = f(v_f)g(T) \tag{17}$$

where $f(v_f)$ is a function only of the free volume (and hence implicitly of the specific volume), or of the liquid configuration, and $g(T)$ is a function of temperature only. For non-polar liquids at least, the value of the latter function changes relatively slowly with T, hence the viscosity is determined primarily by the free volume. Since the internal mobility is also determined by the free volume, the manifestation of the second-order transition as an isoviscous state in *substances of low molecular weight* is consistent with Eq. (17).

The macroscopic viscosity of a polymer depends not only on the free volume and temperature but also on the length of the polymer chain, Z.[3,4] Equation (17) must be replaced therefore by a relationship of the form

$$\eta = f(v_f)g(T)h(Z), \tag{18}$$

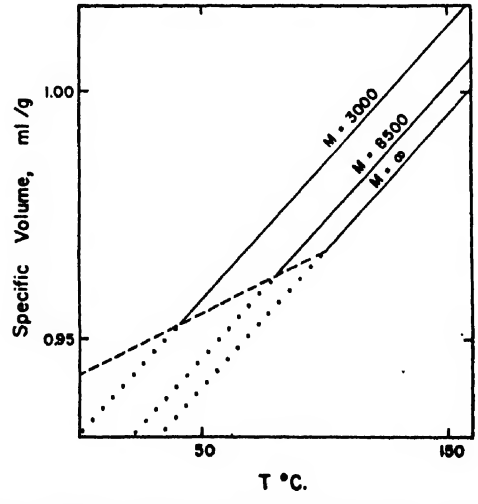


FIG. 7. Specific volume-temperature relations for polystyrene. The parallel lines represent the equilibrium curves, the dotted portions being non-realizable. The dashed line is the v-T curve for polystyrenes in the glassy state.

[20] H. Eyring, J. Chem. Phys. 4, 283 (1936).
[21] Alfrey, Goldfinger, and Mark, J. App. Phys. 14, 700 (1943).
[22] A. J. Batschinsky, Zeits. f. physik. Chemie 84, 643 (1913). D. B. MacLeod, Trans. Faraday Soc. 19, 6 (1923); 21, 151 (1925). W. R. Van Wijk and W. A. Seeder, Physica 4, 1073 (1937); 6, 129 (1939).

where $h(Z)$ is the statistical factor expressing the required coordination of the motions of *all* segments of the polymer molecule. Since the internal mobility depends only on the first two factors (i.e., the rearrangement of local configurational structure does not require long range cooperation of the segments of a given molecule), it is clear that the macroscopic viscosity for a polymer at its second-order transition will vary with M. This is not true of the viscosity-temperature coefficient. If Eq. (18) is rewritten in the form

$$\ln\eta = F(v_f) + G(T) + H(Z) \tag{19}$$

and differentiated with respect to $1/T$, there is obtained

$$\begin{aligned} E_T &= Rd\ln\eta/d(1/T) \\ &= R[(dF(v_f)/dv_f)(dv_f/d(1/T)) + dG(T)/d(1/T) \\ &\quad + (dH(Z)/dZ)(dZ/d(1/T))]. \end{aligned} \tag{20}$$

For non-polar polymers the last two terms on the right should be very small, and presumably may be neglected for the purpose of the present discussion. The quantity $dv_f/d(1/T)$ is independent of M and is not sensitive to temperature [see Eq. (14)]. Hence, E_T depends predominantly on $dF(v_f)/dv_f$, which is a function of the free volume only, and the dependence of E_T on M must be related primarily to the dependence of v_f on M. The constancy of v_f below T_g requires that the first term on the right of Eq. (20) shall be zero below T_g. Hence, the viscosity of polystyrene below T_g will vary only as a result of the contributions of the second and third terms of Eq. (20). To the approximation that these terms may be neglected, the viscosity of a given polystyrene should remain constant at all temperatures below T_g.

No experimental data are available with which to test this latter prediction with regard to polystyrene. Viscosity measurements of silicate glasses, however, have been extended below their second-order transition temperatures. While the temperature coefficient does not fall to zero, it passes through a pronounced maximum at T_g.[23] This behavior is indeed to be expected, when, as in glass, the intermolecular forces between elementary units are large. The second term in Eq. (20) is then appreciable—perhaps comparable in magnitude with the first. The third term must be small or zero.[24] In other words, viscous flow involves a considerable activation, and the resulting temperature dependence persists below T_g notwithstanding the disappearance of the first term in Eq. (20).

Unpublished measurements that we have made of the specific volumes of polyisobutylene fractions at 217° and at lower temperatures indicate that the "free volume" of this polymer is given by

$$v_f = A_f + BT + 32/M \tag{21}$$

which is the analog of Eq. (14) for polystyrene. Here B, the change in the volume-temperature coefficient at T_g, is estimated as $4.5(\pm1)\times10^{-4}$ from the data of Ferry and Parks.[25] If we assume that the second-order transition occurs at a characteristic free volume, then

$$T_g = T_{g,\infty} - 6.9\times10^4/M. \tag{22}$$

Values of T_g calculated from this equation are approximately equal to the corresponding literature values[6,25] provided $T_{g,\infty}$ is taken as $-63°$. This suggests that the correlations observed here for polystyrene may have general validity.

The second-order transition temperature (100°) that we have observed for high molecular weight polystyrene is somewhat higher than values reported elsewhere (Table VI). It is possible that those previously reported may have been vitiated by the presence in the polymer samples of low molecular weight impurities such as unpolymerized monomer, residual solvent, or degradation products formed from the polymer itself.

The volume-temperature coefficients reported here just above and below T_g correspond to values of the expansion coefficient (α) of 6.0 and 2.7×10^{-4}, respectively. These are in good agreement with most of the previously reported values of these coefficients (Table VI). The slightly lower values obtained by some workers for α below T_g may be due to their use of shorter observation times. The lower values obtained in a few cases for α above T_g are unexplained unless, as has been suggested elsewhere,[26] they are due to the presence of impurities. It should be emphasized that the present values were reproducible (within ±10 percent) over wide ranges in temperature for polymers differing widely in molecular weight.

Since the configuration of complex liquids and their volume-temperature dependence are imperfectly understood, we prefer not to hazard an explanation for the non-linearity of the v-T curves above T_g at this time.

SUMMARY

1. The specific volumes of polystyrene fractions ranging in molecular weight from 2970 to 85,000, and of various mixtures of these fractions, have been measured at 217°. The specific volume at a fixed temperature increases linearly with $1/\bar{M}_n$.

2. Measured values of the specific volume-temperature coefficients for these polystyrenes are 7.0, 5.5, and 2.5×10^{-4} above 160°, between the second-order transition temperature (T_g) and 160°, and below T_g, respectively. The last two coefficients appear to be independent of M; the first increases with decreasing M at low molecular weights. Alternately, the results above T_g may be represented by a line with curvature over a wide range in the neighborhood of 160°.

3. The second-order transition temperature is 100 ±2° for high molecular weight polystyrene, and it decreases linearly with $1/\bar{M}_n$.

[23] H. R. Lillie, J. Am. Ceram. Soc. **16**, 619 (1933).
[24] R. W. Douglas, J. Soc. Glass. Tech. **31**, 74 (1947).
[25] J. D. Ferry and G. S. Parks, J. Chem. Phys. **4**, 70 (1936).
[26] F. E. Wiley, Ind. Eng. Chem. **34**, 1052 (1942).

4. At temperatures 10° or more above T_g volume equilibrium appears to coincide with the attainment of temperature equilibrium. In the vicinity of T_g the specific volume decreases approximately linearly with the logarithm of time. In this range the rate of volume relaxation (as measured by $-dv/d\log t$) decreases rapidly with decreasing volume, attaining a low limiting value about 10° below T_g. This limiting value is independent of molecular weight. The rate of internal relaxation (mobility) appears to depend primarily on the internal configurational structure (and on the associated free volume) and to be otherwise independent of T and M.

5. It has been observed that *the specific volumes of polystyrenes below their transition temperatures are independent of the molecular weight.* Hence, it appears that the same "frozen-in" internal configuration prevails in the glassy state independent of molecular weight as well as temperature. The observed dependence of T_g on M can be deduced as a consequence of this conclusion.

6. Bulk viscosities (η) of the same polystyrene samples have been measured from 217 to 75°. The quantity $\log(\eta_T/\eta_{217})$ may be expressed as the product of two factors, one of which depends only on the absolute temperature (as T^{-6}) and the other on $\bar{M}_n$. This ratio at a fixed temperature decreases exponentially with $1/\bar{M}_n$.

7. Generalized relationships for the viscosity of liquids are suggested which indicate that the second-order transition for simple liquids may represent an isoviscous state but that this concept should fail (as it did here) for polymers differing in M. The molecular weight dependence of the viscosity-temperature coefficient for polystyrene above T_g is related logically to its specific volume-molecular weight relationship; the observed non-dependence of this coefficient on M for polystyrenes at their second-order transition temperatures is shown to be a consequence of the non-dependence on M of the specific volumes of these polymers at (or below) their second-order transition temperatures.

8. It is predicted that the viscosity of polystyrene should become relatively independent of T below T_g. The maximum observed in the viscosity-temperature coefficient for a soda-lime-silica glass is explained on the basis of the above relationships.

9. A relationship between T_g and M for polyisobutylene has been suggested.

TABLE VI. Second-order transition temperature and thermal expansion coefficients for polystyrene.

T_g	$\alpha\times10^4$ Above T_g	Below T_g	Reference
≥100°	≥6	0.4 to 2.5	a
100°	6.0	2.7	b
88°	6.47	2.18	c
87°	6.22	2.33	d
82°	6.0	2.6	e
82°	4.63	2.15	f
82°	4.6	2.5	g
82°	—	-	h
—	4.25	2.7	i
- -	4.25	2.4	j
78°	5.89	2.11	k

[a] See reference 5.
[b] Present work.
[c] H. J. Kolb and E. F. Izard, J. App. Phys. 20, 564 (1949).
[d] See reference 2.
[e] W. Patnode and W. J. Scheiber, J. Am. Chem. Soc. 61, 3449 (1939).
[f] F. E. Wiley, Ind. Eng. Chem. 34, 1052 (1942).
[g] See reference 21.
[h] R. F. Boyer and R. S. Spencer, J. App. Phys. 15, 398 (1944).
[i] See reference 13.
[j] R. S. Spencer and G. D. Gilmore, J. App. Phys. 20, 502 (1949).
[k] R. J. Clash and L. M. Rynkiewicz, Ind. Eng. Chem. 36, 279 (1944).

Electrical Breakdown in CSF_8

RONALD GEBALLE AND FRED S. LINN
Department of Physics, University of Washington, Seattle, Washington
(Received December 27, 1949)

The breakdown potential of a new gas, CSF_8, has been measured over the range from $p\delta=4$ to 200 mm×cm under conditions approximating plane-parallel geometry. A comparison of breakdown in air Freon-12 and CSF_8 in the same apparatus indicates the ratio of the strengths of these gases to be approximately 1:2:3, respectively. CSF_8 decomposes rapidly during breakdown into CF_4 and SF_4 with a consequent doubling of pressure and appreciable increase in breakdown potential.

INTRODUCTION

THE insulating properties of certain gaseous compounds of carbon or sulfur and the halogens, notably Freon-12 (CCl_2F_2) and sulfur hexafluoride (SF_6) have been discussed recently by several investigators.[1–5] The two substances named above are found to have breakdown strengths of about 2 to 2.5 times that of air under comparable conditions. It is believed that the high breakdown strengths of such compounds result from the large number of electronegative atoms contained in each molecule, together with the existence of many degrees of freedom capable of absorbing energy from electrons in inelastic collisions. The above two compounds are particularly useful as insulators because they possess, in addition to notable breakdown potentials, the rather large vapor pressures of 5.5 and 22 atmospheres at room temperature, respectively. Recently, in connection with certain syntheses of fluorine-containing compounds, a substance with the empirical formula CSF_8 was isolated by Mr. Gene A. Silvey of the Department of Chemistry. This compound is colorless, and chemically inert. It possesses a boiling point of $-20°C$ at atmospheric pressure and a saturation vapor pressure of 4.3 atmospheres at 20°C. Its structure is believed to be F_3C-SF_5. A detailed description of its synthesis and chemical behavior will be discussed elsewhere.[6] In view of the large number of fluorine atoms per molecule and the high vapor pressure of CSF_8 a survey of its insulating properties was felt to be of considerable interest and was undertaken by the present authors. Measurements of the breakdown strength of this gas were made in the range from $p\delta=4$ to 200 mm×cm under conditions closely approximating plane-parallel geometry. A comparison of its strength with those of air and Freon-12 was made under the same conditions. It was found that air, Freon-12, and CSF_8 possess strengths in the approximate ratio 1:2:3 in this range of $p\delta$. It was also found that CSF_8 decomposes in an unusual manner in the course of electrical breakdown. The breakdown products could be separated in relatively pure states and their identity verified.

APPARATUS

The breakdown voltage was measured in a Pyrex discharge vessel containing two plane Monel disks of 7.65 cm diameter and spaced 1.5 cm apart. Because of the limited supply of CSF_8 the volume of the envelope was deliberately made small, about 0.8 liter. To localize discharges on the axis of the electrode system under this condition a hemispherical boss of 1.6 mm radius was mounted at the center of the anode. The small field distortion thus introduced was responsible, in certain cases, as will be seen, for the production of streamer pulses preceding breakdown. The distortion was sufficiently small, however, that no serious alteration in breakdown potential was noticed in standard gases in the range of $p\delta$ under investigation. In computing $p\delta$, the electrode separation, neglecting the boss, was used.

The pumping system consisted of a mercury diffusion pump backed by a mechanical pump. Pressures were determined by the use of appropriate McLeod gauges or a mercury manometer. Traps surrounded by liquid nitrogen were used to eliminate contamination by mercury vapor except during short intervals when pressure measurements were being made.

Direct voltage up to 40 kv was supplied by a half-wave rectifier with sufficient filtering to reduce ripple to 0.1 percent. Small increments in voltage were made possible by the use of a small transformer in series with the high voltage one, each with its own variable input. The voltage was read with the aid of a calibrated multiplier and a potentiometer. The over-all error in voltage readings, due primarily to line voltage fluctuation, amounted to about one percent. A two-megohm resistor was in series with the high voltage to prevent excessive current upon breakdown.

To insure an adequate supply of initiating electrons a radium-containing sample was placed near the discharge chamber.

PROCEDURE

The electrodes were polished and cleaned before assembly of the system performed in an atmosphere of nitrogen. The tube was annealed in hydrogen and baked

[1] H. C. Pollock and F. S. Cooper, Phys. Rev. **56**, 170 (1939).
[2] Trump, Safford, and Cloud, Trans. A.I.E.E. **60**, 132 (1941).
[3] H. H. Skilling and W. C. Brenner, Trans. A.I.E.E. **61**, 191 (1942).
[4] B. Hochberg and E. Sandberg, J. Tech. Phys. U.S.S.R. **12**, 65 (1942); E. Sandberg, J. Tech. Phys. U.S.S.R. **17**, 299 (1947).
[5] Schumb, Trump, and Priest, Ind. Eng. Chem. **41**, 1348 (1949).
[6] Gene A. Silvey and George H. Cady, J. Am. Chem. Soc. (to be published).

out at 400°C while at a pressure of 10^{-5} mm of mercury. Between runs in different gases the tube was well pumped out. In one case, after the run in Freon-12, the system was completely dismantled and cleaned before proceeding.

The air used in this investigation was withdrawn slowly from the room and passed through a trap surrounded by liquid nitrogen. For all the measurements in this gas, except the last, the vapor pressure of oxygen at the temperature of the trap exceeded the total pressure in the apparatus; hence the composition of the air was not appreciably altered. In any case, the breakdown potentials of air and nitrogen are very closely the same in this range of $p\delta$.[3] The more readily condensed gases were admitted from traps after careful drying by redistillation. Hydrogen, used to test the geometry, was obtained from a commercial cylinder, passed through a catalytic purifier and then through a trap immersed in liquid nitrogen.

Breakdown voltages were found by slowly increasing the applied potential in increments of a few volts, with suitable delay between each step in an attempt to eliminate errors due to statistical factors. In general the onset of a visible discharge was taken as the breakdown potential. In certain cases, as described below, observations were made with an oscilliscope connected across a series resistor.

EXPERIMENTAL RESULTS

1. Hydrogen

The breakdown potential of hydrogen was investigated in order to determine the properties of the particular geometrical arrangement employed. The oscilliscope showed small pulses, increasing in size with increased voltage, which suddenly assumed the character of breakdown streamer pulses.[7] A few volts above this the discharge became visible, first pulsating and then steady to the eye. These events took place at well-defined potentials, reproducible within the accuracy of the apparatus. In Fig. 1 the streamer onset voltage thus found is compared with the curve compiled by Druyvesteyn and Penning[8] for breakdown between plane electrodes. All pressures have been reduced to standard conditions. It can be seen that the advent of the large pulses coincided rather closely with the breakdown voltage found under conditions of strictly uniform field.

2. Freon-12

Commercial Freon-12 was used after vapor density measurements indicated its purity was better than 99.5 percent. The oscilliscope gave no hint of impending breakdown in this gas, but merely exhibited preliminary burst pulses as reported by Mohr and Weissler.[9] In contrast to these workers no hysteresis was noted. There were, however, large fluctuations in delay between the time of application of voltage and breakdown when the voltage was applied by closing a switch. The average of these fluctuations is a function of the applied voltage. It decreases, for example, from 30 sec. at 29,700 volts to 2 sec. at 30,900 volts for $p\delta = 275$ mm×cm. Figure 1 shows the breakdown potentials associated with a maximum delay of 1 min. For comparison the curves of Trump, Safford, and Cloud[2] as well as Hochberg and Sandberg[4] are shown. The decomposition products of Freon-12 were found to be exceedingly corrosive, attacking glass, electrodes, and mercury.

3. Air

The breakdown potential of air was determined for use as a reference. The appearance of visible discharge was regarded as breakdown. There was no noticeable fluctuation in delay times and voltages were easily reproducible. No study of the pre-breakdown pulses was made. The data obtained are compared in Fig. 1 with the curve compiled by Whitehead[10] for plane geometry. They agree with the results of Fisher[11] in the range where comparison is possible.

4. CSF_8

No streamer pulses were observed in this gas. The small burst pulses, which in many gases precede streamers, increased in size and frequency until a spark occurred, persisting until the high voltage switch was

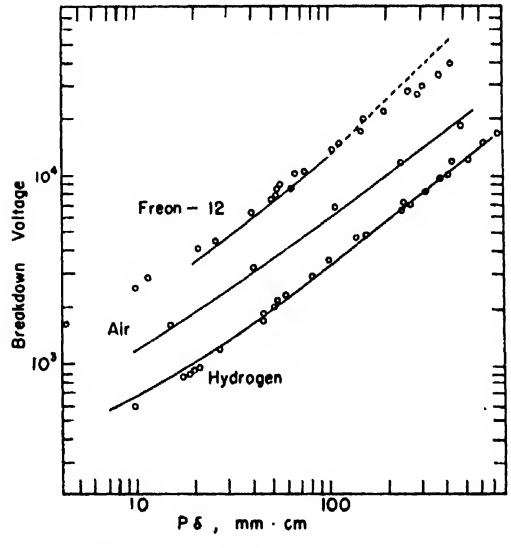


FIG. 1. Breakdown voltages measured in the present chamber as compared with published data. Curves are as follows: H_2, reference 7; air, reference 9; Freon-12, solid portion, reference 4; Freon-12, dashed portion, reference 2. Points are those of the present work. All data reduced to 0°C.

[7] A. F. Kip, Phys. Rev. **55**, 549 (1939).
[8] M. J. Druyvesteyn and F. M. Penning, Rev. Mod. Phys. **12** (April, 1940).
[9] B. A. Mohr and G. L. Weissler, Phys. Rev. **72**, 289 (1947).
[10] S. Whitehead, *Dielectric Phenomena* (D. Van Nostrand Company, Inc., New York, 1927), p. 43.
[11] L. H. Fisher, Ph.D. thesis, Library, University of California, Berkeley, California.

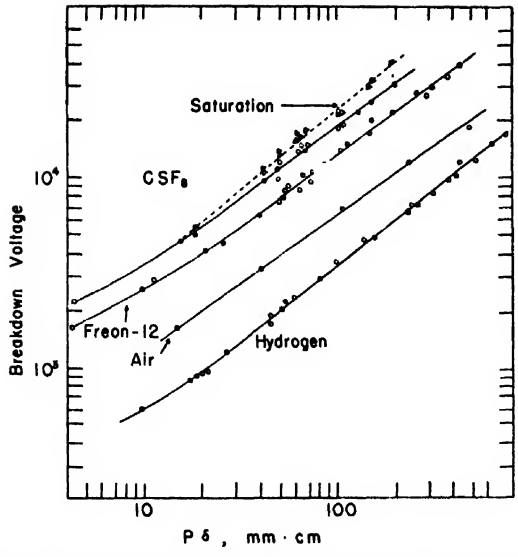


FIG. 2. Comparison of breakdown voltages of CSF_8 with other gases. Curves are drawn through the experimental points. Dashed curve is the saturation voltage discussed in the text. All data reduced to 0°C.

opened. On closing the switch again, the potential remaining constant, breakdown did not recur. When the voltage was increased to a new level it could be reapplied several times to produce sparking. The time lapse between the closing of the switch and the spark increased for each of these breakdowns from an initial value of less than a second to more than five minutes and eventually it was apparent that no more sparks were possible at this level. At higher levels the sequence of events was similar except that the total number of breakdowns at each voltage increased. In addition, the time lag increased less rapidly with the number of breakdowns at a given level. At a level about ten percent above the initial breakdown potential a saturation voltage was attained after which sparking took place with small time lags which fluctuated between zero and about three seconds without steady increase.

It was believed that an investigation of the decomposition products would provide the explanation of these observations. To obtain the products in sufficient quantity for analysis, a continuous discharge of 2 ma was permitted to pass through CSF_8 for 15 min. and the gases collected over liquid nitrogen. This process was repeated several times. During the course of these runs measurements of the pressure indicated an increase to twice the initial value, after which no further change occurred. Analysis of the products by Mr. Silvey provided the following information:

(1) When condensed, the products formed two immiscible phases.

(2) One of the phases was definitely identified through its vapor density and freezing point as CF_4.

(3) The other constituent, because its odor, melting and boiling points agreed closely with those reported by Fischer and Jaenchner,[12] was tentatively identified as SF_4, a compound of hitherto dubious existence.[5]

(4) The average vapor density of the products was correct assuming the above are the sole products.

It appears then that during the course of electrical discharge the CSF_8 is split rapidly into two products, CF_4 and SF_4. The consequent rise in pressure accounts for the behavior of the breakdown potential in the following manner: When the first breakdown took place some of the CSF_8 was decomposed, the pressure rose and no further breakdown was possible at this voltage. When the voltage was increased and reapplied, breakdown occurred immediately, producing further decomposition. At a constant pressure the breakdown potential of the products is considerably less than that of the CSF_8. Hence, because of the slow diffusion rate of the heavy molecules involved, the mixture of gases between the electrodes broke down when the same voltage was again applied, even though the pressure had increased to some extent. With successive applications of voltage the decomposition continued, the mixture of gases became more nearly homogeneous until the pressure increased to such an extent that no further breakdown could take place at this potential. At higher voltages the process was similar. Because the percentage of CSF_8 continually diminished, the lag introduced by diffusion became less pronounced until with complete decomposition the saturation voltage was attained. It should be noted that the lags after reaching saturation were not dependent on voltage in the pronounced manner of Freon-12. To verify the nature of the decomposition products an additional sample of CSF_8 was subjected to discharge. Again the doubling of pressure and rise in breakdown potential were noted. Separation of the two phases and analysis for sulfur and fluorine by Mr. Silvey indicated without doubt that one of the products was SF_4.

The breakdown potential of CSF_8 as well as the saturation voltage characteristic of the completely decomposed gas under conditions of constant volume are shown in Fig. 2 with pressures reduced to 0°C. For comparison, the results obtained under Sections 1–3 are given. It can be seen that in the range of the present investigation the strengths of air, Freon-12, and CSF_8 are in the approximate ratio 1:2:3.

At the conclusion of the discharges in CSF_8 a thin whitish layer, identified as sulfur, coated the glass. The Monel electrodes used in the first set of measurements showed no evidence of corrosion. The vessel used to decompose the gas in the second set contained stainless steel, which became markedly corroded. These effects are probably due to further decomposition during the passage of the current. When allowed to stand for a day or so, the mixture of CF_4 and SF_4 exhibited a constant pressure and breakdown potential.

It is a pleasure to thank Professor George H. Cady and Mr. Silvey of the Department of Chemistry for supplying the CSF_8 and to acknowledge the aid rendered by the latter in performing the various analyses.

[12] J. Fischer and W. Jaenchner, Zeits. f. angew. Chemie 42, 810 (1929).

The Effect of Cold-Work Distortion on X-Ray Patterns*

B. E. Warren and B. L. Averbach
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received January 9, 1950)

With modern experimental technique, it is possible to measure a peak shape with sufficient accuracy to justify an interpretation based on the precise shape of the reflection. The corrected shape is represented by a cosine Fourier series and a set of A_n coefficients determined. A plot of the A_n coefficients *vs.* n will distinguish between distortion and particle size broadening. From the A_n coefficients, root mean square values of strain averaged over lengths na_3 are obtained. The decrease in these values for increasing length na_3 is a direct indication of the non-uniform nature of the strains in cold-worked metal. By measuring several orders of a given plane, it is theoretically possible to obtain a distribution function of the strains directly from a Fourier transform of the A_n coefficients.

I. INTRODUCTION

COLD-WORK distortion produces a broadening of powder pattern reflections. Most of the previous work in this field has involved measurements of cold-work broadening as a function of λ and θ, and the use of these breadth measurements to distinguish between various assumed theories of the nature of cold-work distortion.

With modern experimental technique, using focusing curved crystal monochromators and Geiger counter spectrometers, it is possible to measure the shape of a powder pattern line with sufficient accuracy to justify an interpretation based on the precise shape of the line, rather than using just the line breadth.[1] By making full use of the shape of the line, it is possible to get a great deal of information about the nature of cold-work distortion without making any *a priori* assumptions. The theoretical treatment must be carried through for a completely general distortion, so that what we get in the end is given uniquely by the x-ray measurements and is not partly the result of assumptions concerning the nature of the distortion which have been inserted during the treatment.

II. EFFECT OF A GENERAL DISTORTION

We start with the same formulation used by Stokes and Wilson.[2] Only cubic crystals are considered for which it is always possible to adopt a set of orthogonal axes so as to make any reflection of the form $(00l)$. The position of any cell $m_1m_2m_3$ is given by the vector

$$R_m = m_1a_1 + m_2a_2 + m_3a_3 + \delta_{m_1m_2m_3},$$

where δ is an arbitrary displacement, in general different for every cell. Expressed in electron units the intensity from one crystal is

$$I = F^2 \sum_{m_1} \sum \sum \sum \sum \sum_{m_3'} \exp[2\pi i/\lambda (s - s_0)\cdot(R_m - R_{m'})],$$

where s_0 and s are unit vectors giving the directions of the primary and diffracted beams. Let

$$s - s_0 = \lambda(h_1b_1 + h_2b_2 + h_3b_3),$$

where $b_1b_2b_3$ are the reciprocal vectors, and $h_1h_2h_3$ are continuous variables. The intensity per crystal is

$$I(h_1h_2h_3) = F^2 \sum \sum \sum \sum \sum \sum_{m_3'} \exp\Big[2\pi i \; h_1(m_1 - m_1') + h_2(m_2 - m_2') + h_3(m_3 - m_3') + \left(\frac{s - s_0}{\lambda}\right)\cdot(\delta_m - \delta_{m'}) \Big\} \Big]. \quad (1)$$

For a powder pattern, the total diffracted power in a reflection is given by an integration throughout a sufficient volume in the reciprocal lattice to include everything contributing to the reflection.

$$P = \frac{MjR^2\lambda^3}{4v_a} \int\int\int \frac{I(h_1h_2h_3)}{\sin\theta} dh_1dh_2dh_3, \quad (2)$$

where M is the number of crystals in the sample, j is the hkl multiplicity, R is the sample to film distance, and v_a is the volume of the unit cell. The distribution of power with angle is expressed by $P_{2\theta}$ where

$$P = \int P_{2\theta} d(2\theta). \quad (3)$$

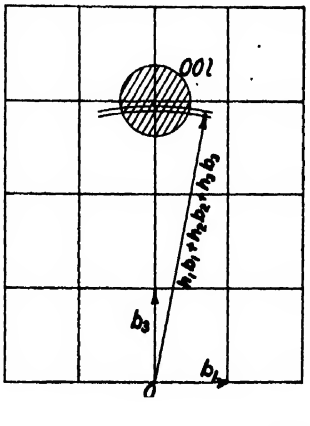


FIG. 1. Approximation of a spherical surface by a tangent plane involved in the integration about 00*l*.

* Research sponsored by the ONR under Contract N5 ori-07832.
[1] B. L. Averbach and B. E. Warren, J. App. Phys. **20**, 885 (1949).
[2] A. R. Stokes and A. J. C. Wilson, Proc. Camb. Phil. Soc., **38**, 313 (1942); Proc. Phys. Soc. **56**, 174 (1949).

In carrying out the integration about the point $00l$ in Fig. 1, it is assumed that the spreading out of the point due to distortion is small compared to its distance from the origin, so that a spherical layer can be approximated by a horizontal layer

$$|s-s_0| = 2\sin\theta = \lambda|h_1b_1+h_2b_2+h_3b_3| \rightarrow \lambda|b_3|h_3$$

$$dh_3 = \frac{\cos\theta d(2\theta)}{\lambda|b_3|}.$$

The integration with respect to dh_3 in (2) can be expressed in terms of $d(2\theta)$ and from (3) we get $P_{2\theta}$ by leaving out this integration.

$$P_{2\theta} = \frac{MjR^2\lambda^2}{4v_a|b_3|\tan\theta}\int\int I(h_1h_2h_3)dh_1dh_2. \qquad (4)$$

Experimentally we measure the distribution of power per unit length of diffraction line

$$P_{2\theta}' = \frac{P_{2\theta}}{2\pi R\sin 2\theta}.$$

Let

$$K = -\frac{MjR\lambda^2F^2}{16\pi v_a|b_3|\sin^2\theta}.$$

The distribution of power is then

$$P_{2\theta}' = K\int\int \sum_{m_1}\sum\sum\sum\sum\sum_{m_3'} \exp\left[2\pi i\left\{h_1(m_1-m_1') + h_2(m_2-m_2') + h_3(m_3-m_3') + \left(\frac{s-s_0}{\lambda}\right)\cdot(\delta_m-\delta_{m'})\right\}\right]dh_1dh_2. \qquad (5)$$

Represent the displacement δ_m by

$$\delta_m = X_ma_1 + Y_ma_2 + Z_ma_3 \quad (m = m_1, m_2, m_3)$$

and for computing the scalar product $(s-s_0)\cdot(\delta_m-\delta_m')$ use the average position of the vector $(s-s_0)$, $\langle s-s_0\rangle_{Av} = \lambda h_3b_3$

$$\frac{2\pi i}{\lambda}(s-s_0)\cdot(\delta_m-\delta_{m'}) = 2\pi i h_3(Z_m-Z_{m'}).$$

It is only the component of displacement perpendicular to the reflecting planes that is involved. To include everything belonging to point $00l$ we carry out the integration with respect to dh_1 and dh_2 from $-\frac{1}{2}$ to $+\frac{1}{2}$.

$$P_{2\theta}' = K\sum_{m_1}\sum\sum\sum\sum\sum\sum_{m_3'} \frac{\sin\pi(m_1-m_1')}{\pi(m_1-m_1')} \times \frac{\sin\pi(m_2-m_2')}{\pi(m_2-m_2')} \exp[2\pi i h_3\{(m_3-m_3') + (Z_m-Z_{m'})\}]. \qquad (6)$$

The sine functions are zero except for $m_1=m_1'$ and $m_2=m_2'$. This means that we should think of the crystal in terms of columns of unit cells, the columns being parallel to a_3 or perpendicular to the reflecting planes $00l$. The summation with respect to m_3 and m_3' is carried out for all pairs of cells in a given column, and the summation with respect to m_1 and m_2 means summation of the contributions from each column.

Let

$$n = m_3 - m_3' \quad Z_n = Z_{m3} - Z_{m3}'.$$

Assuming that the number of cells in a column N_3 is large compared to the value of n at which a summation with respect to n will converge

$$\sum_{m_3}\sum_{m_3'} \rightarrow N_3\sum_{n=-\infty}^{+\infty}$$

$$P_{2\theta}' = K\sum_c N_3\sum_{n=-\infty}^{+\infty}\langle\exp[2\pi i h_3(n+Z_n)]\rangle_{Av}, \qquad (7)$$

where $\sum_c$ means summation over the columns, and the average is an average over-all pairs of cells $m_1m_2m_3$ and $m_1'm_2'm_3'$ with $m_1=m_1'$, $m_2=m_2'$ and $m_3-m_3'=n$. Let the total number of cells in the crystal be $N=\sum_c N_3$. Since $Z_{-n}=-Z_n$

$$P_{2\theta}' = KN\sum_{n=-\infty}^{+\infty}\langle\cos 2\pi h_3(n+Z_n)\rangle_{Av}$$

$$P_{2\theta}' = KN\sum_{n=-\infty}^{+\infty}(A_n\cos 2\pi nh_3 + B_n\sin 2\pi nh_3), \qquad (8)$$

where $A_n = \langle\cos 2\pi h_3Z_n\rangle_{Av}$ and $B_n = -\langle\sin 2\pi h_3Z_n\rangle_{Av}$.

The coefficients of the sin terms will vanish if positive and the negative values of Z_n are equally probable for a given n. If they do not vanish their effect is to displace the center of the $P_{2\theta}'$ peak from the position for an undistorted material. This is just another way of recognizing that cold-work distortion may produce an average expansion or contraction of the lattice. Measurements on cold-worked metals to which the theory has been applied so far, show no appreciable shift in peak position, and hence we will confine our attention to the restricted case where the sin terms are negligible.

$$P_{2\theta}' = KN\sum_{n=-\infty}^{+\infty}A_n\cos 2\pi nh_3, \qquad (9)$$

where

$$h_3 = 2a_3(\sin\theta)/\lambda$$
$$A_n = \langle\cos 2\pi lZ_n\rangle_{Av}.$$

In the expression for A_n we have replaced h_3 by its peak value l. From (9) the corrected shape of the peak is represented by a cosine Fourier series, and the coefficients A_n depend only on relative displacements perpendicular to the reflecting planes. By definition a_3Z_n is the change in length of a column of original length na_3.

III. EXPERIMENTAL DETERMINATION OF FOURIER COEFFICIENTS

The coefficients A_n are determined from carefully measured peak shapes from a cold-worked sample and from an annealed sample. The second peak is used to correct the first for all sources of broadening other than that due to cold-work distortion. The correction is carried out by the method of Stokes[3] using the Lipson-Beevers strips. This method delivers directly the desired A_n coefficients. The Fourier analysis is done over the range h_3' from $-\frac{1}{2}$ to $+\frac{1}{2}$ where $h_3'=h_3-l$. Fictitious values of l and a_3 are used where necessary to make the range of the Lipson-Beevers strips cover efficiently the range of 2θ-values in the measured peak.

IV. EFFECT OF PARTICLE SIZE BROADENING

Particle size broadening is represented by a cosine Fourier series similar to that developed for distortion broadening and hence the Fourier coefficients A_n give a very general method of handling either effect. The particle size broadening follows the same treatment used for distortion broadening up to Eq. (6) except that δ and hence Z_m and $Z_{m'}$ are zero.

$$P_{2\theta}'=K\sum_{m_1}\sum\sum\sum\sum\sum_{m_3'}\frac{\sin\pi(m_1-m_1')}{\pi(m_1-m_1')}\times\frac{\sin\pi(m_2-m_2')}{\pi(m_2-m_2')}\exp[2\pi i h_3(m_3-m_3')].$$

Consider the crystal in terms of columns of unit cells perpendicular to the reflecting planes. Let n_i be the number of columns of length i cells, so that

$$N=\sum_{i=1}^{\infty} in_i$$

is the total number of cells in the sample.

$$P_{2\theta}'=KN\sum_{n=-\infty}^{+\infty}A_n\cos2\pi nh_3, \tag{10}$$

where

$$A_n=\frac{1}{N}\sum_{i=|n|+1}^{\infty}(i-|n|)n_i.$$

Introduce a distribution function $p(i)$ such that $p(i)di$ is the number of columns of length between i and $i+di$ cells.

$$\int_1^{\infty}p(i)di=N_c \qquad \int_1^{\infty}ip(i)di=N,$$

where N_c is the number of columns and N is the number of cells. In terms of the distribution function

$$A_n=\frac{1}{N}\int_{i=|n|+1}^{\infty}(i-|n|)p(i)di.$$

[3] A. R. Stokes, Proc. Phys. Soc. **61**, 382 (1948).

For both distortion and particle size broadening the coefficients A_n have been defined so that A_0 is unity.

$$\frac{dA_n}{dn}=-\frac{1}{N}\int_{i=|n|+1}^{\infty}p(i)di.$$

The initial slope of the A_n *vs.* n curve is

$$(dA_n/dn)_{n=0}=-N_c/N=-1/\bar{N}_3, \tag{11}$$

where $\bar{N}_3a_3$ is the average column length and hence an average particle size in the direction a_3.

$$(d^2A_n)/(dn^2)=(1/N)p(n). \tag{12}$$

If the measurements of particle size broadening are expressed in terms of a plot of the Fourier coefficients A_n *vs.* n the initial slope of the curve gives directly the average column length and the second derivative gives the column length distribution. This important result has already been obtained by M. F. Bertaut[4] by essentially the same method. The treatment given here is one developed by the authors before they were aware of Bertaut's work and it has been included here so as to present the particle size theory in the same notation used for distortion broadening.

V. REPRESENTATION BY FOURIER COEFFICIENTS

The representation of measured peak shapes by a plot of the Fourier coefficients A_n *vs.* n gives a very general and powerful method for representing many different effects in the sample. Three different examples are shown in Fig. 2. For a distortion such as that of cold-work the coefficients are given by

$$A_n=\langle\cos2\pi lZ_n\rangle_{Av}.$$

For small n, we expect the values of $\langle Z_n^2\rangle_{Av}$[4] to be small and proportional to n and hence A_n *vs.* n should start with the shape of an error curve. The effect of small particle size is to give a finite initial slope whose value is determined by the average particle size. This

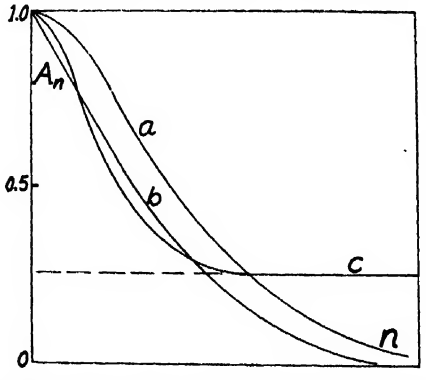


FIG. 2. Difference in shape of the A_n *vs.* n curves. (a) Cold-work distortion. (b) Particle size broadening. (c) Temperature effect.

[4] M. F. Bertaut, Comptes Rendus **228**, 492 (1949).

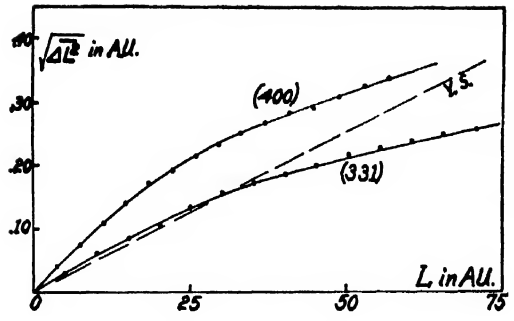


FIG. 3. Relative displacements for cold-worked 70–30 α-brass. ΔL is the change in length of a column of length L, both expressed in angstrom units.

expected difference in the shape of the A_n *vs.* n curves has been confirmed by the experimental measurements which have been made. From precise measurements of the shape of a *single reflection* it is possible to distinguish between particle size and distortion broadening.

The effect of temperature vibration is represented schematically by curve c. For large n, the value of $\langle Z_n^2\rangle_{Av}^{\frac{1}{2}}$ becomes independent of n, and A_n approaches a constant value. If each A_n value is split into the part above and below the horizontal line of curve c, the complete set of values below the line represents the sharp reflection unaffected in breadth but cut down in height, while that part of the coefficients above the line represents a diffuse scattering which peaks at the position of the reflection.

The integral breadth β of a reflection is given by

$$\beta=\frac{1}{(P_{2\theta}')_{\max}}\int P_{2\theta}'d(2\theta)$$

$$\beta=\frac{1}{\sum_{-\infty}^{+\infty}A_n}\int_{-\frac{1}{2}}^{+\frac{1}{2}}\sum A_n\cos(2\pi nh_3)\frac{\lambda}{a_3\cos\theta}dh_3$$

$$\beta=\frac{\lambda}{a_3\cos\theta}\frac{1}{\sum_{-\infty}^{+\infty}A_n}. \qquad (13)$$

From a measurement of a peak breadth β we get only the sum of the Fourier coefficients $\sum A_n$, and hence not nearly as much information as is available in a knowledge of the value of each coefficient.

VI. INTERPRETATION OF MEASUREMENTS ON COLD-WORKED METAL

From the experimental peak shape we determine a set of coefficients A_n which involve the relative displacements Z_n. If the values of Z_n are sufficiently small, or if they follow a Gaussian distribution $p(Z_n)=C\times\exp -\alpha^2Z_n^2$, the average of the cosine can be expressed as an exponential.

$$A_n=\langle\cos 2\pi lZ_n\rangle_{Av}\rightarrow\exp[-2\pi^2l^2\langle Z_n^2\rangle_{Av}]. \qquad (14)$$

From each coefficient A_n we compute a root mean square relative displacement $\langle Z_n^2\rangle_{Av}^{\frac{1}{2}}$.

Figure 3 shows a set of measurements on cold-worked 70–30 α-brass filings. To avoid the confusion resulting from the fictitious a_3 values used in the Fourier analysis, the Z_n values computed by (14) have been changed to distances in angstrom units.† L is a column length and ΔL is the change in length of the column due to cold-work

$$L=na_3 \quad \Delta L=a_3Z_n.$$

The slope of a line from the origin to any point on a curve represents a root mean square strain averaged over columns of length $L=na_3$ perpendicular to the reflecting planes. The initial slopes represent the true root mean square strains for the respective directions. The initial slopes are roughly in the inverse ratio to the values of Young's modulus for the corresponding directions, in agreement with previous studies involving measurements of peak width only.[5,6] For the cold-work involved in preparing filings, the initial slopes are of the same order as the yield strain computed from the yield stress and an average value of Young's modulus.

If strains were constant over appreciable distances in the material, as is usually assumed in the micro stress theory of cold-work broadening, the curves of Fig. 3 should be straight lines. To the extent that the approximation in (14) is justified, the noticeable drooping of these curves at distances as small as 20A is evidence for the non-uniform and localized nature of the strains in cold-worked metal.

The curves of Fig. 3 depend on the approximation that the Z_n values are either small or represented by a Gaussian distribution function $p(Z_n)=C\exp[-\alpha^2Z_n^2]$. In principle it is possible to obtain the distribution function directly from the x-ray data rather than assuming it. By using different wave-lengths it is possible to throw several orders of a particular set of planes into the high angle region. By measuring the A_n coefficients for several values of l and making use of the fact that $A_n=1.0$ for $l=0$, and vanishes for large l, we determine $A_n(l)$ approximately as a function of the variable l. In terms of the distribution function

$$A_n(l)=\langle\cos 2\pi lZ_n\rangle_{Av}=\int_0^\infty p(Z_n)\cos 2\pi lZ_ndZ_n.$$

The distribution function is given directly as the

† In a previous presentation of similar curves (reference 1) the scale of ordinates for the 331 and 222 curves involved fictitious a_3 values different from the cube edge a and the slopes are not comparable.

[5] C. S. Smith and E. E. Stickley, Phys. Rev. **64**, 191 (1943).

[6] Stokes, Pascoe, and Lipson, Nature **151**, 137 (1943).

Fourier transform of the measured $A_n(l)$ values.

$$p(Z_n)=4\int_0^\infty A_n(l)\cos 2\pi Z_n l dl. \qquad (15)$$

In principle we can determine an average strain distribution function for each n and plane (hkl). The plane (hkl) selects a direction in the crystal perpendicular to the planes, the index n selects a distance $L=na_3$ along this direction, and $p(Z_n)$ gives the distribution of strain magnitudes averaged over such lengths na_3. This probably represents as much information about the nature of cold-work deformation in a metal as can be obtained directly from the x-ray broadening data, with no assumptions about the nature of the strains being introduced.

A Quantized Theory of Strain Hardening as Applied to the Cutting of Metals

MILTON C. SHAW
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received January 18, 1950)

A theory of strain hardening is presented which utilizes the fact that metals deform blockwise rather than continuously. A metal is assumed to possess an orderly array of weak spots through which slip planes pass. The slip displacement which occurs upon a single plane is related to the spacing of weak points and the slip plane inclination. Use of a linear relation between shear plane displacement and resisting shear stress, enables the mean shear stress to be computed.

In the two examples considered it is seen that the influence of normal stress on the shear plane is less important in cutting than has been generally believed. Strain hardening and the short range inhomogeneity of metals is seen to account for most of the dynamic increase in shear strength during cutting, including the increase in shear energy per unit volume with decreased depth of cut. This size effect is akin to the increase in tensile strength with decreased specimen diameter, both phenomena involving strain hardening as influenced by inhomogeneity. The quantity Aa^2 which represents in a single number the strain hardening and short range inhomogeneity characteristics of a metal should prove particularly valuable in analyzing the cutting characteristics of metals which strain harden extensively.

INTRODUCTION

FUNDAMENTAL studies in the field of solid state physics have revealed that, in general, metals do not strain uniformly. When a single metal crystal is loaded, it will first deform elastically and then plastically. The plastic deformation is observed to occur by a process of slip as one section of the crystal slides over another. The direction of this shear deformation in a single crystal is along those planes containing the greatest number of atoms per unit area. Slip planes are clearly evident under the microscope on a polished and etched single crystal that has been plastically deformed. The presence of slip planes of a fairly regular spacing attests the fact that the metal has not been strained uniformly throughout but rather that the deformation is confined to well-defined planes that are many atomic layers apart. It is further found that polycrystalline metals also strain blockwise rather than uniformly.

Energy considerations show that slip does not involve the motion of one layer of uniformly spaced atoms over another en masse. Present theory rather suggests that slip occurs due to the motion of certain types of lattice imperfections called dislocations through a crystal. Dislocations are generally associated with highly localized stress gradients in a metal which may arise from any one of a number of causes (small internal cracks or holes, the mutual interference of adjacent grains at grain boundaries or small inclusions of foreign matter). While the exact physical nature of dislocations is not exactly clear, the properties of dislocations explain a number of characteristics of metals that are not as conveniently explained in other ways. For the purpose of the present discussion the dislocation sources may be considered as weak spots which are present in a piece of metal that has been processed in a particular way. The planes along which slip occurs are those which pass through one or more of these weak spots, the metal between adjacent planes being undeformed.

FIG. 1. Photomicrograph of a partially formed chip. Magnification 100×.

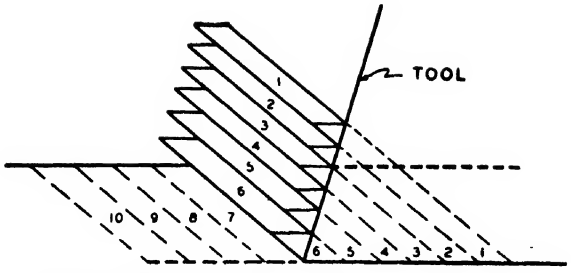


FIG. 2. Schematic representation of cutting process.

Despite the foregoing inhomogeneous picture of plastic deformation, it has been customary to treat the metal as a continuum in quantitative calculations of strain. Consideration of the well-known phenomenon of strain hardening makes it evident that a quantized treatment of strain should be used in place of a treatment involving uniformly distributed strain. The metal cutting process is a convenient one for illustrating the need for adopting such a discontinuous picture of the straining process.

THE METAL CUTTING PROCESS

When metal is cut by a two-dimensional single point cutting tool the material removed is plastically deformed to a considerable extent. Photomicrographs of the chip-tool interface of a partially formed chip, such as that shown in Fig. 1, show that metal is removed in a cutting operation by a process of shear. A distinct line of demarcation in structure (AB) is generally evident in such photomicrographs, below which metal is undeformed plastically and above which it has a deformed structure. The angle which AB, the shear line, makes with the direction of cut is called the shear angle, ϕ.

Piispanen[1] first likened the metal cutting process to the successive displacement of a series of cards as shown in Fig. 2. In this figure portions of the chip numbered 1, 2, 3 etc. originally occupied the similarly numbered positions on the workpiece. In practice the thickness of the individual "cards" is quite small, corresponding to the spacing of the aforementioned weak spots in a piece of metal. The mechanism of cutting shown exaggerated in Fig. 2 forms the basis for the theory of metal cutting that is presented here.

Previous attempts to formulate a quantitative theory of metal cutting have ignored the important tendency of a metal to strain harden. Piispanen[1] has mentioned that metals strain harden and that this factor should be included in a complete analysis of the cutting process. Merchant[2] has also mentioned that strain hardening may occur in cutting but concludes that it plays an insignificant role. Drucker[3] has more recently discussed the possibility of strain hardening in metal cutting and concludes that metal behaves as an ideal plastic when cut, strain hardening not being among the items of interest.

To prove conclusively that strain hardening can play a very significant role in metal cutting, one need only attempt to machine a specimen of austenitic stainless steel. Herbert[4] has presented data for such materials which clearly show the importance of strain hardening when they are machined. The source of the conclusion that strain hardening plays an insignificant role in metal cutting arises from the classical view that strain in a plastically deformed metal is uniformly distributed rather than being confined to certain planes. This matter will be considered in greater detail in the following discussion.

PRESENT THEORY OF METAL CUTTING

Before discussing the manner in which strain hardening may be introduced into cutting calculations, it may be well to review briefly the present theory of metal cutting that has been developed for the most part by Piispanen and Merchant. This theory is for cutting operations which produce a continuous chip with no built-up edge. The forces acting at the point of a two-dimensional cutting tool are shown in Fig. 3. If the chip above the shear plane (shown heavy) is taken as a free body, then forces R and R' must be equal if the negligible inertia force due to the change in velocity of the metal as it flows across the shear plane is ignored. Force R' is shown resolved into two sets of force components F_H and F_V, those which are normally measured with a cutting dynamometer; and F_S and N_S, those along and perpendicular to the shear plane. Force R is shown resolved into the friction and normal components on the tool face, F and N. Attention is drawn to angles α

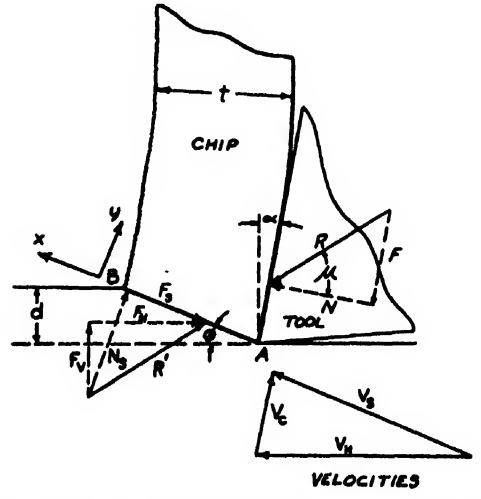


FIG. 3. Force and velocity relations in a two-dimensional cutting operation.

[1] V. Piispanen, "Lastunmuodostumisen Teoriaa" (Theory of Chip Formation), Teknillinen Aikakauslehti 27, No. 9, 315–322 (1937).
[2] M. E. Merchant, J. App. Phys. 16, 267–275, 318–324 (1945).
[3] D. C. Drucker, J. App. Phys. 20, 1013–1021 (1949).
[4] E. G. Herbert, "Work Hardening Properties of Metals," Trans. ASME 48, 705–748 (1926).

(the rake angle), ϕ (the shear angle), and μ (the friction angle) which are defined in the figure. A velocity triangle is also shown in Fig. 3 where V_H is the absolute velocity of the tool, V_S is the velocity of the chip along the shear plane relative to the stationary work piece, and V_C is the velocity of the chip up the face of the tool.

For purposes of analysis, the shear and normal stresses may be considered constant across the shear plane although actually it is evident that this cannot be the case.[5] We may also ignore the fact that plastic deformation in the vicinity of the tool point may not be along the main shear plane but at a much greater angle of shear as is evident in Fig. 1.

Piispanen[1] and Merchant[2] have suggested that the shear stress to cause slip along the shear plane is influenced by the normal stress on the shear plane and given to a good approximation by the expression

$$\tau = \tau_0 + B\sigma \tag{1}$$

where τ_0 is the critical shear stress to cause slip in the absence of normal stress, σ is the normal stress on the shear plane, and B is a constant. Under certain conditions a dependence of critical shear stress or normal stress has been observed experimentally by Bridgman[6] for polycrystalline metals.

Both Piispanen[1] and Merchant[2] have applied the principle of minimum energy to the cutting process in order to relate the shear angle to other known variables. From the geometry of Fig. 3, it is evident that for a cut of unit width

$$F_H V_H = \tau d V_H \frac{\cos(\mu-\alpha)}{\sin\phi\cos(\phi+\mu-\alpha)} \tag{2}$$

where $F_H V_H$ is the rate at which work is expended in the cutting process. Equation (1) may be written

$$\tau = \tau_0 + B\tau\tan(\phi+\mu-\alpha) \tag{3}$$

and by combining Eqs. (2) and (3)

$$F_H V_H = \tau_0 d V_H \cdot \times \frac{\cos(\mu-\alpha)}{\sin\phi\cos(\phi+\tau-\alpha) - B\sin\phi\cos(\phi+\tau-\alpha)}. \tag{4}$$

Applying the principle of minimum energy by differentiating with respect to ϕ (considering μ constant[7]) and equating to zero gives

$$\cot(2\phi+\mu-\alpha) = B. \tag{5}$$

[5] In order that the shear plane correspond to a plane of maximum shear stress, it is necessary that at each point along the shear plane a compressive stress exist that is directed parallel to the shear plane and which is equal to the compressive stress acting normal to the shear plane. The compressive stress parallel to the shear plane will be zero at the free surface of the chip and have its maximum value at the end of the shear plane adjacent to the tool point.

[6] P. W. Bridgman, J. App. Phys. 14, 273–283 (1943).

[7] Drucker (see reference 3) has pointed out that in this step μ has been considered independent of ϕ but it is not at all clear that this is so.

This equation which relates the shear, friction, and rake angles to the slope of the shear stress-normal stress curve (B) will be discussed later.

STRAIN HARDENING

As already mentioned when metal is cut at a temperature sufficiently below its melting point, it shears in blocks rather than uniformly. Under ordinary conditions, the shear planes will be very closely spaced corresponding to the closeness of spacing of the weak spots in the metal. It may be assumed that slip planes are so spaced that a single weak spot is present on each plane. Drucker[3] has recently employed a random array of weak points to qualitatively demonstrate the increase in unit cutting energy with decrease in depth of cut. However, inasmuch as the spacing of weak

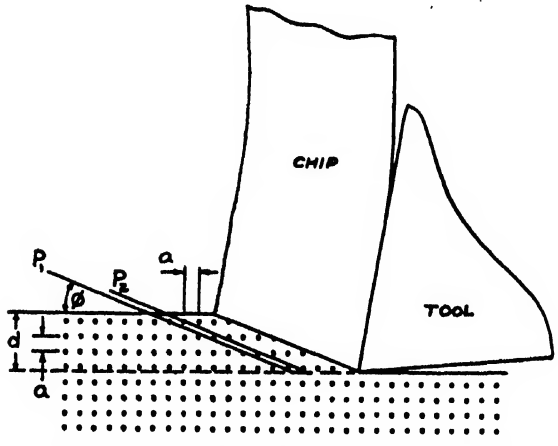


FIG. 4. Specimen with uniform array of weak spots.

points is very small compared with usual depths of cut, an orderly array of weak spots seems justified. The dots shown in Fig. 4 represent such an orderly arrangement of weak points to an exaggerated scale. These points have a uniform spacing of a units in each direction.

Let P_1 and P_2 be two shear planes making an angle ϕ with the direction of cut and passing through adjacent points in the first row below the surface. If the depth of cut is d, then d/a planes may be placed between those at P_1 and P_2 such that a single plane passes through each weak spot in the layer in the process of being cut. The number of planes per inch in a direction perpendicular to the shear plane will be

$$n = \frac{d}{a(a\sin\phi)} \tag{6}$$

or the spacing of successive planes is

$$y = \frac{a^2\sin\phi}{d}. \tag{7}$$

The total slip on a given shear plane will be

$$x=\frac{a^2\sin\phi}{d}\gamma \tag{8}$$

where γ is the unit uniform strain which from Figs. 1 and 3 may be shown to be

$$\gamma=\cot\phi+\tan(\phi-\alpha). \tag{9}$$

Need now exists for data concerning the manner in which shear stress is related to shear strain. As a guide to the nature of a relation which might be assumed the so-called true stress-strain curves for metals tested in direct tension may be considered. A representative curve of this type is shown in Fig. 5. Such curves are characterized by a linear elastic region from 0 to P_1 and a less steep nearly linear plastic region extending from P_2 to the breaking point P_3. For purposes of analysis involving large strains (as in metal cutting), the curve shown dotted may be substituted for the actual diagram. Thus, according to this approximation, a critical stress σ_0 exists below which there is no plastic flow, and the stress required to produce a given strain increases linearly with the strain. The slope of this curve represents the best known measure of the tendency of a metal to strain harden.

A diagram similar to that of Fig. 5 is obtained if shear stress is plotted against shear strain. Such a diagram is shown in Fig. 6 with total shear strain x as the abscissa rather than unit strain. The critical shear stress to initiate slip is τ_0' while the slope of the curve is represented by the symbol A in the following discussion. Point P_3 is the point at which slip on a given plane ceases. Following Piispanen and Merchant we may assume that τ_0' is influenced linearly by normal stress as follows

$$\tau_0'=\tau_0+B\sigma. \tag{10}$$

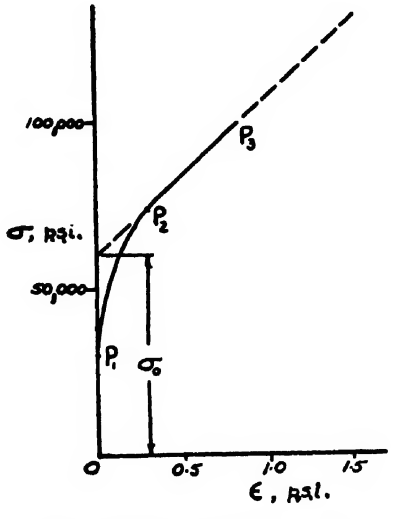


FIG. 5. True stress-strain diagram.

It is then evident that the mean shear stress during slip is

$$\tau=\tau_0+B\sigma+(Ax/2) \tag{11}$$

or from Eqs. (8) and (9)

$$\tau=\tau_0+B\sigma+\frac{Aa^2\sin\phi}{2d}[\cot\phi+\tan(\phi-\alpha)]. \tag{12}$$

This equation differs from that employed by Piispanen and Merchant by the inclusion of the last term on the right which introduces the effect of strain hardening and the inhomogeneity of the metal cut. It is interesting to note that the slope of the stress-strain curve (A) and the spacing of weak points (a) appear in the equation only in the form of the product Aa^2. This product may thus be considered a new constant of the material which measures its tendency to strain harden during shear flow. For a given material the strain hardening effect is seen to increase linearly with the unit uniform strain γ which the chip undergoes (the quantity in brackets in Eq. (12)), and inversely with the first power of the depth of cut d.

The most convenient method of evaluating the constants τ_0, Aa^2, and B for a given metal is from cutting data obtained by use of a tool dynamometer. While theoretically the data from but three different cuts would be sufficient to enable these constants to be determined, actually it is advisable to use from 10 to 20 points so that probability may tend to neutralize the effect of the natural scatter in the data due to long range in homogeneity of metal cut and other causes.

FIRST EXAMPLE

In order to illustrate the method of evaluating the three constants from cutting tests, some data previously published by Merchant[2] will be considered. In these tests a particular specimen of NE9445 steel having a Brinell hardness of 187 was cut with a carbide tool. The workpiece was in the form of a tube so as to give a simple two-dimensional cut when the tool was fed axially against the end of the tube. The chips produced were in the form of a continuous ribbon, the condition on which the foregoing analysis is based. The data obtained for a range of cutting speed, depth of cut and rake angle are summarized in Table I. Merchant plotted these data in accordance with Eq. (1) to obtain Fig. 7. The values of τ and B obtained from this plot were 58,000 lb./in.^{2} and 0.23 respectively, corresponding to the solid line.

These same data may be replotted in accordance with Eq. (12) developed here. For this purpose it is convenient to rearrange the equation as follows

$$\tau-B\sigma=Aa^2[\cot\phi+\tan(\phi-\alpha)]\frac{\sin\phi}{2d}+\tau_0$$

$$=Aa^2\left(\frac{\gamma\sin\phi}{2d}\right)+\tau_0, \tag{13}$$

where it is evident that a straight line of slope Aa^2 and intercept τ_0 will be obtained when the quantity $(\tau - B\sigma)$ is plotted against $(\gamma \sin\phi)/2d$. Such a plot could be readily obtained from the data of Table I if B were known. A procedure which may be followed in the absence of this value is to assume successive values of B and obtain a series of plots. The line best fitting the data points corresponds to the value of B to be adopted. A series of plots obtained in this manner is shown in Fig. 8 for values of B of 0, 0.05, 0.1, 0.2 and Merchant's value of 0.23 which of course should be a horizontal line. The foregoing procedure of plotting a series of curves involves less work than it might appear at first glance due to the ease of obtaining successive curves from the one first plotted.

It would appear from Fig. 8 that the curve best fitting the data points corresponds to a B of 0.05 and the corresponding equation for shear stress may be written

$$\tau = 80{,}000 + 0.05\sigma + 29\frac{\gamma \sin\phi}{2d}, \text{ p.s.i.} \qquad (14)$$

In considering which value of B best fits the data points, special attention should be given those points corresponding to high and low depths of cut. Data corresponding to a greater range of depth of cut will enable a more decisive choice of the correct value of B. The value of B in Eq. (14) is seen to be far less than that in Merchant's analysis and in better agreement with the value of 0.125 directly observed by Bridgman[6] for drill rod. It would thus appear that much of the effect which Merchant has attributed to normal stress on the shear plane is actually due to strain hardening and the inhomogeneity of the metal cut.

Merchant[2] has plotted directly measured values of ϕ against the quantity $(\mu - \alpha)$ from Eq. (5) in which B is taken equal to 0.23. This plot corresponds to the solid line in Fig. 9. If Eq. (12) including the effect of strain hardening is substituted for Eq. (1) in the derivation leading to Eq. (5) it is found that for minimum energy

TABLE I. Data for NE9445 steel.[2]

Depth of cut in. ×1000	Cutting speed V_H ft./min.	Rake angle α, deg.	ϕ deg.	γ	tan μ	σ psi ×10⁻³	τ psi ×10⁻³	Ws in. lb./in.³ ×10⁻³
3.70	197	10	17.0	3.4	1.05	115	85	292
3.70	400	10	19.0	3.1	1.11	137	88	266
3.70	642	10	21.5	2.7	0.95	129	90	248
3.70	1186	10	25.0	2.4	0.81	130	93	224
3.70	400	−10	16.5	3.9	0.64	153	89	342
3.70	637	−10	19.0	3.5	0.58	152	90	311
3.70	1160	−10	22.0	3.1	0.51	154	94	289
1.09	542	10	19.0	3.1	1.12	164	103	308
2.34	542	10	18.5	3.1	1.08	140	92	287
3.70	542	10	21.5	2.7	0.96	131	91	248
7.88	542	10	25.0	2.4	0.76	116	89	214
1.09	542	−10	12.5	5.0	0.78	179	103	523
2.34	542	−10	16.0	4.0	0.70	172	96	385
3.70	542	−10	19.0	3.5	0.60	164	94	306
7.88	542	−10	22.5	3.1	0.46	138	90	270

(again considering μ to be independent of ϕ)

$$\cot[\cot^{-1}(B) - \phi - \mu + \alpha] = \cot\phi - \frac{\tan(\phi-\alpha)\tan\phi\sin(2\phi-\alpha)}{\cos\alpha\left(1+\frac{2d\tau_0}{Aa^2}\right)}. \qquad (15)$$

This is seen to reduce to Eq. (5) if strain hardening is ignored (i.e., when $Aa^2 = 0$). For most applications the last term on the right side of Eq. (15) is negligible and Eq. (15) reduces to Eq. (5). The strain hardening term will be important only when the quantity $(d\tau_0)/(Aa^2)$ is small, i.e., if Aa^2 is large, the depth of cut is very small as in grinding, or the material has a low critical shear stress τ_0.

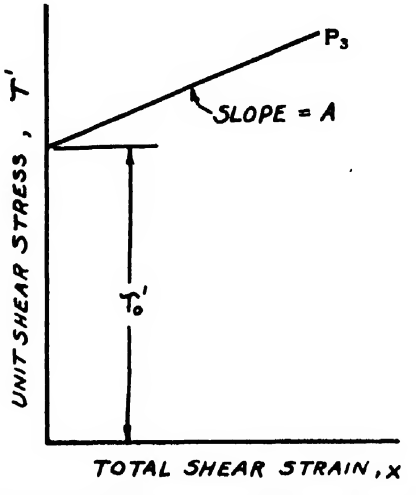


FIG. 6. Unit shear stress-total shear strain diagram.

Equation (15) has been used to obtain the dotted curve shown in Fig. 9. In this instance the second term

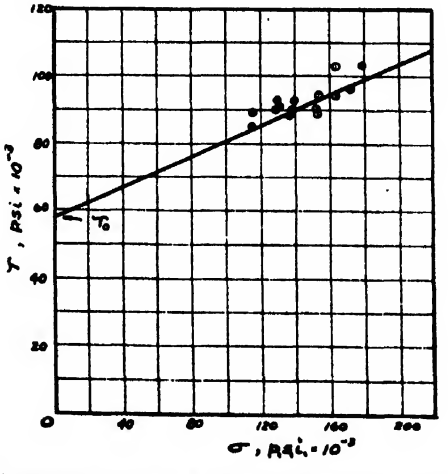


FIG. 7. Relation between shear strength and normal stress for cutting data of Table I.

on the right is negligible. It is evident that the dotted curve does not fit the data as well as the solid curve but this observation is not thought very significant in view of the following approximations of the theory leading to Eqs. (5) and (15).

1. The influence of temperature upon τ_0 has been ignored. This effect is apt to be quite significant since in practice the mean temperature along the shear plane is quite high. The effect of temperature is to cause a decrease in the value of ϕ obtained from Eqs. (5) or (15). The inclusion of the temperature effect would thus tend to lower both curves of Fig. 9, bringing the dotted line into better agreement with the experimental points but the solid line into poorer agreement.

2. The coefficient of friction has been considered independent of the shear angle in the foregoing analysis. It is not known how significant this approximation is but it may be observed that a decrease in the coefficient of friction that accompanies the use of an effective cutting fluid causes an appreciable increase in the angle ϕ. Conversely, it is to be expected that a change in ϕ will be accompanied by a change in μ. This point

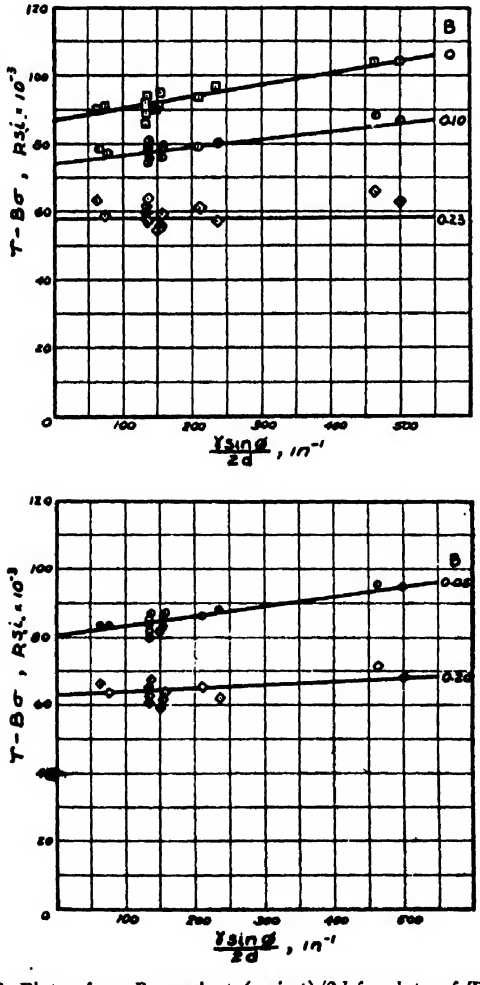


FIG. 8. Plots of $\tau - B\sigma$ against $(\gamma \sin\phi)/2d$ for data of Table I.

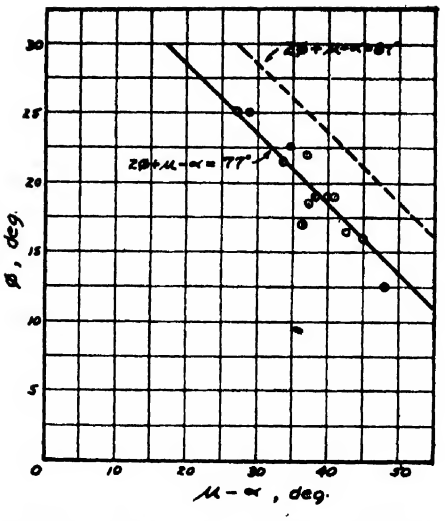


FIG. 9. Plot of ϕ against $\mu - \alpha$ for data of Table I.

places the portion of the analysis leading to Eqs. (5) and (15) in question.

3. The effect upon τ_0 of the change in the rate of strain with ϕ has been ignored in the derivation of Eqs. (5) and (15). If this effect were included, it would also cause the dotted and solid curves of Fig. 9 to be lower. However, it is unlikely that this effect would be significant inasmuch as the rate of strain is always at a very high level in metal cutting. As is evident in Fig. 1, the metal goes from an essentially unstrained condition to the fully strained condition in the very short time it takes to traverse the shear plane. Theoretically this would correspond to an infinite rate of strain were the shear plane a truly mathematical surface of zero thickness.

The position of the solid curve in Fig. 9 is significantly influenced by the choice of curve in Fig. 7. In this instance it would appear that the points in Fig. 7 call for a line of considerably less slope which would tend to shift the solid line of Fig. 9 in the direction of the dotted line.

SECOND EXAMPLE

Results of a series of tests were recently published[8] in which seamless tubing of 6 inch diameter and 0.475 inch wall thickness was cut using two-dimensional high speed steel tools having rake angles ranging from 25 to 45°. The steel was cut under conditions to prevent the formation of a built-up edge at a speed of 90 feet per minute. It is interesting to examine these data which are summarized in Table II from the point of view of the foregoing theory.

In accordance with Merchant's theory, the relation of shear strength to normal stress is first considered (Fig. 10). In this instance it is quite difficult to know

[8] Lapsley, Grassi, and Thomsen, "Correlation of Plastic Deformation during Metal Cutting with Tensile Properties of the Work Material, ASME Preprint No. 49-A-121 (December, 1949).

just where to draw the line, there being apparently two possibilities—the line shown solid and that shown dotted. The data are next shown plotted in Fig. 11 in accordance with the strain hardening theory presented here. Only the curve corresponding to $B=0$ is shown inasmuch as other values of B investigated gave points which fitted a straight line less closely than that for $B=0$. The solid line shown corresponds to the equation

$$\tau = 52{,}000 + 153(\gamma \sin\phi)/2d, \text{ p.s.i.} \qquad (16)$$

In the third figure of this series (Fig. 12) the observed values of ϕ are shown plotted against the quantity $(\mu-\alpha)$. Equation (15) again reduces to Eq. (5) since the strain hardening term is negligible. The relation based on strain hardening theory is given by the dash line in Fig. 12 while the relation predicted by Merchant's theory (corresponding to the solid line in Fig. 10) is shown solid. If the dotted line in Fig. 10 is used, the dotted line of Fig. 12 is obtained. The line in Fig. 12 corresponding to strain hardening theory is seen to be in good agreement with the data for tools of high rake angle, but the data for the 25° rake angle tool lies below the curve as in the example of the previous section and probably for the same reasons

DISCUSSION

Drucker[3] has considered the fact that essentially all of the strain which occurs in the cutting process takes place as the material crosses the shear plane at approximately a constant shear stress to be a paradoxical result. He concludes that this can only mean that the metal behaves as an ideal plastic which strains without strain hardening ($A=0$). This is a natural conclusion when viewed from the classical point of view where the metal is considered to strain uniformly. However, while there is no appreciable spatial change in shear stress in the vicinity of the shear plane, there is a significant temporal

TABLE II. Data for mild steel.*

Depth of cut in. ×1000	Cutting speed V_H ft./min.	Rake angle α, deg.	ϕ deg.	γ	tan μ	σ psi ×10⁻³	τ psi ×10⁻³	W_s in. lb./in.³ ×10⁻³
2.5	90	25	20.9	2.55	1.46	104	82.8	209
3.5	90	25	21.5	2.48	1.46	96.0	74.6	180
5	90	25	24.0	2.23	1.375	100	75.5	169
6	90	25	20.1	2.65	1.36	86.5	65.8	175
8.5	90	25	22.4	2.38	1.38	83.6	66.5	159
2.5	90	35	31.6	1.56	1.53	96.0	71.4	112
3.5	90	35	31.9	1.55	1.52	84.0	61.2	96.4
5	90	35	32.0	1.55	1.48	82.9	62.4	96.5
6	90	35	32.2	1.54	1.425	88.6	61.8	95.5
8.5	90	35	32.0	1.55	1.38	80.9	58.5	91.3
2.5	90	40	38.7	1.32	1.54	94.5	72.0	94.0
3.5	90	40	35.4	1.33	1.50	84.3	66.5	87.8
5	90	40	37.5	1.26	1.44	87.0	66.1	83.1
6	90	40	37.2	1.27	1.42	90.0	63.5	81.1
8.5	90	40	37.2	1.27	1.35	83.1	61.5	78.1
2.5	90	45	41.9	1.06	1.83	114.6	71.1	74.8
3.5	90	45	41.9	1.06	1.74	98.8	64.2	68.5
5	90	45	40.2	1.10	1.64	87.2	63.8	69.6
6	90	45	39.6	1.01	1.60	80.8	61.7	62.7
8.5	90	45	39.9	1.11	1.51	73.6	58.7	65.0

* See reference 8.

change. The shear stress on the shear plane is not constant with time but increases as slip proceeds due to strain hardening. This variation in shear stress would give rise to a variable shearing force which could be detected with a very sensitive dynamometer of high frequency response were it not for the fact that the process of slip is in varying phases of completion at different points across the chip. This may be seen by examining the back surface of a chip where the edges of certain slip planes are evident as wavy rather than straight lines extending across the chip in a direction perpendicular to the direction of cut. The slip process is thus not quite as simple as the sliding of cards one over the other, but the actual process appears to involve the propagation of screw type as well as edge type dislocations.

The stress concentration associated with a sharp tool point is largely responsible for the fact that strain occurs essentially on only one plane at a time. In this

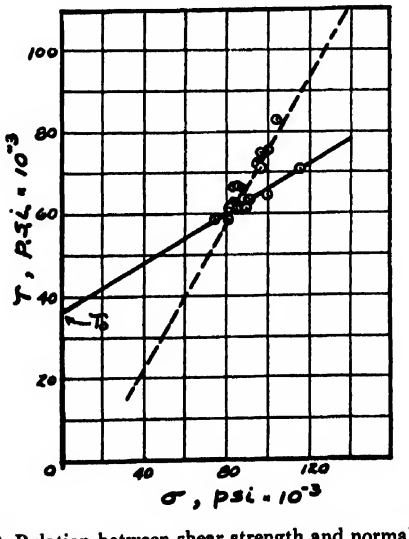


FIG. 10. Relation between shear strength and normal stress for cutting data of Table II.

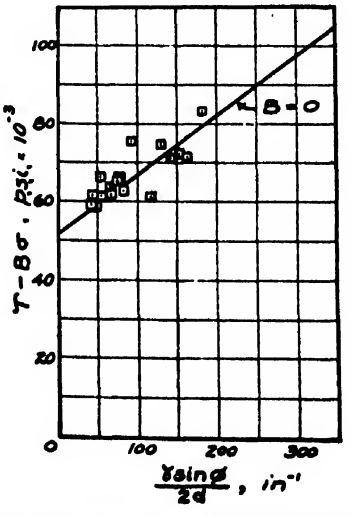


FIG. 11. Plot of $\tau - B\sigma$ against $(\gamma \sin\phi)/2d$ for data of Table II.

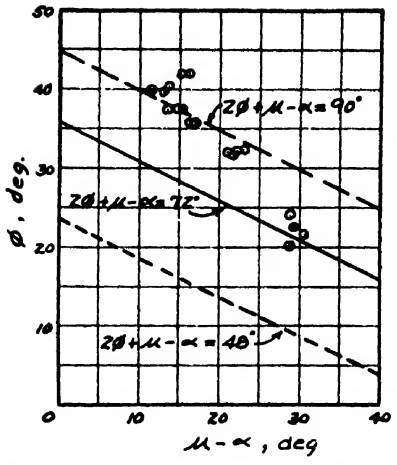


FIG. 12. Plot of ϕ against $\mu-\alpha$ for data of Table II.

respect the cutting process differs from all others in the field of plastic flow. For example, in the ordinary tensile test, strain occurs simultaneously upon many planes as the load is progressively increased. The concept of a homogeneous strain is not so serious in such applications as in the case of the cutting process. However, it would appear that even in such problems it is advantageous to adopt the quantized picture of strain which will enable a size effect to be included in the quantitative theory just as the effect of the depth of cut was included in the foregoing theory of metal cutting.

In Table II it is evident that the shear work to remove a cubic inch of metal increases with decreased depth of cut. This is particularly true for the data obtained with the 25° rake angle tool. The strain hardening theory explains this observation. The shear work for this metal may be obtained by use of Eq. (16) as follows

$$W_S = \tau\gamma = 52{,}000\gamma + 153(\gamma^2 \sin\phi)/2d, \text{ in.-lb./in.}^3. \quad (17)$$

From this expression it is evident that the specific shear work is increased with an increase of shear strain γ or a decrease of depth of cut d. The experimental values of W_S given in Table II are in agreement not only with the observation concerning the influence of depth of cut but also shear strain. The specific work is seen to increase as the rake angle α is decreased, the main reason for this being the increase in strain associated with the decrease in rake angle.

CONCLUSIONS

As the basic metal cutting process becomes better understood, the gap between our well-known materials tests and the metal cutting operation is narrowed. The simple two dimensional machining operation is seen to provide means for conducting materials tests that are now carried out only with considerable difficulty. Specifically the cutting test offers a very convenient method for determining the dynamic shear strength of metals since the inherent feature of this process is to allow shear on approximately one plane at a time thus naturally providing very high rates of strain. Correspondingly high rates of strain are obtained in the ordinary tensile test only with considerable difficulty and uncertainty due to inertia effects. Furthermore the characteristics of the cutting process enable tests to be carried to strains that are considerably higher than those in the usual tensile test. The simplicity of the test apparatus and the fact that a specially shaped specimen need not be prepared are additional factors of interest. While the tensile test has a superficial appearance of simplicity, a little study will reveal that this process may even at present be less completely understood than the cutting process. It would therefore not be surprising to one day find the simple two-dimensional cutting process in use in materials laboratories engaged not alone in machinability studies but also in use in evaluating the basic physical properties of metals.

Mechanical and Electrical Properties of Plasticized Vinyl Chloride Compositions*

LAWRENCE E. NIELSEN, ROLF BUCHDAHL, AND RITA LEVREAULT
Plastics Division, Monsanto Chemical Company, Springfield, Massachusetts

(Received January 24, 1950)

The temperature dependence of the dynamic mechanical and a.c. electrical properties of a given polyvinyl chloride plasticizer composition can be characterized by two quantities: (a) the transition temperature which is defined as the temperature of maximum dissipation factor and (b) the half-widths of the loss factor-temperature curve. A simple relationship exists between the transition temperature determined mechanically or electrically and the volume fraction of a compatible plasticizer. The modification of this relationship for non-compatible plasticizer has been indicated. The half-width is a function of the volume fraction of the plasticizer, although this relationship is not as simple as the one obtained for the transition temperature. A simple correlation exists between the electrical and mechanical half-width. The compliance measured in tensile creep has been found to be a function of the applied load and the compliance-load relationship varies depending on the region in which the material finds itself. The shape of the creep curve at different temperatures has been correlated with the dissipation factor-temperature curve. The shape of the stress-strain curve has been related to the modulus-temperature curve and a correlation between the initial Young's modulus and the ultimate elongation and breaking strength has been established. It has been shown that the relaxation theory is not sufficient to explain certain features of the modulus (or dissipation factor) temperature relationship. The limited significance of the μ-value and of the viscosity of plasticizers with respect to plasticizer efficiency has been demonstrated.

INTRODUCTION

THE mechanical properties of plasticizer-polymer systems in the solid state have been the subject of several investigations.[1–4] As a result of such work, it has been possible to draw the following conclusions concerning the effect of plasticizer concentration and type of plasticizer on these properties: (1) the addition of plasticizer always lowers and broadens the transition or softening region compared to the unplasticized resin; (2) plasticizers differ in their effectiveness to lower and broaden the transition region.** Several attempts have been made to interpret these data in terms of other physical quantities. For example, Jones[5] tried to show that the viscosity and the viscosity temperature coefficient of the plasticizer are significant quantities determining plasticizer efficiency, or the transition range of plasticizer-polymer systems. On the other hand, Boyer and Spencer[6] have derived an expression for the flex or second-order transition temperature as a function of polymer concentration, assuming that this temperature corresponds to an isoelastic or isoviscous state of this system, and second, making use of the Flory equation of viscous flow. The same authors have also suggested[6] that the reason for the difference in kind of plasticizers might be due to the difference in μ-value for different plasticizers in various polymers. The importance of plasticizer-resin interaction has also been pointed out by Aiken[4] in interpreting the very different creep curves which one obtains for polyvinyl chloride-plasticizer systems. Alfrey[7] has tried to show that the creep data support in general his relaxation theory of viscoelastic deformations of linear high polymers. However, the experimental data available until now are neither extensive nor precise enough to permit a proper evaluation of the proposed mechanism of plasticizer action. It was, therefore, one of the major objectives of this work to undertake a systematic study of various mechanical properties, such as the dynamic modulus and dissipation factor, stress-strain relationships and creep, over a wide temperature range as a function of plasticizer concentration and type of plasticizer, in order to obtain more definite information concerning the dependence of these properties on the above-stated variables. In Table I are shown the various plasticizers used, and it will be noted that all of them, with the exception of di-lauryl phthalate, are commonly referred to as solvent-type plasticizers; such generalizations, which will be permissible on the basis of the results obtained, apply, therefore, primarily to this type of plasticizer, although some characteristic effects of non-solvent type plasticizers will be considered. In order to study the significance of the μ-value on plasticizer action, without changing the functional groups of the molecule, five different phthalate plasticizers were chosen, covering a μ-value range in polyvinyl chloride from -0.1 to $+1.4$.[8]

The a.c. electrical properties of plasticizer-polymer systems have been studied by several investigators;[9–11]

* Presented in parts at the American Physical Society Meeting, New York City (January, 1949) and at the High Polymer Forum of the A.C.S. Atlantic City Meeting (September, 1949).

[1] R. Lawrence and E. B. McIntyre, Ind. Eng. Chem. 41, 689 (1949).

[2] N. C. Read and Harding, Ind. Eng. Chem. 41, 675 (1949).

[3] Rider, Sumner, and Myers, Ind. Eng. Chem. 41, 709 (1949).

[4] Aiken, Alfrey, Janssen, and Mark, J. Polymer Sci. 2, 178 (1947).

** The transition region has, in the past, been characterized by some arbitrary temperatures, such as the flex or brittleness temperature or the apparent second-order transition temperature; the significance of the transition region lies in the fact that it is in this range where all mechanical properties undergo a considerable change in magnitude and kind over a rather limited temperature range.

[5] H. Jones, J. Soc. Chem. Ind. 67, 415 (1948).

[6] R. F. Boyer and Spencer, J. Polymer Sci. 2, 157 (1947).

[7] T. Alfrey, Jr., *Mechanical Behavior of High Polymers* (Interscience Publishers, Inc., New York, 1948).

[8] W. G. Bagley (unpublished data).

[9] Davies, Miller, and Busse, J. Am. Chem. Soc. 63, 361 (1941).

[10] Würstlin, Kolloid Zeits. 105, 9 (1943).

[11] R. Fuoss, J. Am. Chem. Soc. 61, 2334 (1939); 63, 2401, 2410, 2832 (1941); 64, 283 (1942).

TABLE I.

1	2	3	4	5	6	7	8	9	10	11
		Concentration		Transition temp. °C		Half-width °C				
Composition	Density	c_v	c_m	Mech.	Elec.	Mech.	Elec.	$\left(\frac{E_t}{G''}\right)_{25°C}$	ΔE (kcal.)	30°C $\eta(c.p)$
Polyvinyl chloride	1.41[a]	0	0	86	105	7(?)	15–21	2.09	—	—
PVC-dimethyl phthalate	1.19	0.291	0.10	28	42	13	38	—	14.3	9.8
PVC-diethyl phthalate	1.12	0.254	0.07	35	50	13	40	2.09	7.6	10.0
PVC-DEP		0.336	0.10	22	32	17(?)	45	1.24	15.2	—
PVC-DEP		0.443	0.15	−6	15	17	46	2.67	26.9	—
PVC-dibutyl phthalate	1.045	0.158	0.03	57	70	12	26	1.93	—	10.4
PVC-DBP		0.313	0.07	25	39	42	—	2.68	9.6	—
PVC-DBP		0.400	0.10	4	21	13	45–51	2.14	8.4	—
PVC-DBP		0.526	0.15	−20	3	12(?)	45	1.59	14.9	11.5
PVC-*n*-dioctyl phthalate	0.986	0.402	0.07	5	21	46(?)	61	2.62	—	—
PVC-*n*-DOP		0.501	0.10	−15	3	32	74	2.22	12.6	—
PVC-dilauryl phthalate	0.943	0.368		—	68	—	85	—	—	—
PVC-tricresyl phosphate	1.16	0.351	0.069	29	45	15(?)	37	—	15.5	15.4
PVC-TCP		0.443	0.10	12	30	15	42	2.86	34.0	—
PVC-dibutyl sebacate	0.935	0.250	0.042	34	46	56	20	1.87	—	7.6
PVC-DBS		0.458	0.10	−27	−7	62	31	1.52	—	—

[a] Density values are for resin and plasticizer and not for the mixture.

the work of Fuoss,[11] in particular, has given us extensive data concerning the dependence of the dielectric constant and the power factor on temperature, on plasticizer concentration and type of plasticizer. However, the theoretical interpretation of such data on the basis of a simple relaxation mechanism has not been too successful.[12] It was hoped, therefore, that a simultaneous investigation of the electrical and mechanical properties *on identical samples* might aid in the clarification of the plasticizer mechanism in general, particularly as a close qualitative correlation between the electrical and mechanical properties of plasticizer systems had been suggested in the past.[13]

EXPERIMENTAL DETAILS

Preparation and Composition of Specimens

A polyvinyl chloride of fine particle size, having an intrinsic viscosity of about one in cyclohexanol, was used in the preparation of all specimens. The volume and mole fractions of all the compositions used and the density of the resin and the plasticizers are given in Table I, columns 3, 4, and 5, respectively. The preparation of all test specimens was carried out according to the following procedure. The resin, the plasticizer, and a small amount of tribase stabilizer (two percent based on the amount of the resin used) were weighed out separately and mixed together by hand. The mix was then transferred to a two-roll mill and was milled for not longer than 10 min. at a temperature varying between 265° and 280°F. The mix was taken off the mill, cut into small pieces, and specimens of various thickness were compression-molded to the required dimensions using, as much as possible, the same molding cycle throughout.

Instruments and Testing Procedures

The dynamic mechanical properties were measured using a torsion pendulum, with the exception of a few measurements on unplasticized polyvinyl chloride, which were carried out with a vibrating reed-type instrument.[14] The rectangular test strip ($2\times\frac{3}{8}\times0.025$ in.) is part of the pendulum, see Fig. 1. The dynamic modulus is calculated from the frequency of the free vibration and the dimensions of the specimens, and the damping is determined from the logarithmic decrement of the vibration. Actually, the frequency and the damping are measured as a function of the load attached to the lower end of the specimen and all data given are extrapolated to zero load. The temperature of the instrument is controlled by an air bath, having a temperature variation of about ±0.5°C. The frequency of the vibration can be varied by a factor of 10, using different disks giving the system different moments of inertia. The precision of the shear modulus and logarithmic decrement values is ±5 and ±10 percent, respectively, provided the damping is not too high; as the damping increases the precision of the measurement decreases.

The tensile creep was measured by clamping both ends of the strip, applying a fixed load to the lower clamp and following the separation of two inked lines on the sample with a cathetometer as a function of time. The sample, but not the load, was surrounded by an air bath, controlled within ±0.5°C. A difference of 0.005 cm in length could be measured with the cathetometer. In general, the creep data could be reproduced within ±5 percent, except in the transition region, provided a new sample was used in the repeat run. It was found that the previous history definitely affected the absolute value of the creep and the shape of the creep

[12] F. G. Kirkwood and R. Fuoss, J. Chem. Phys. 9, 329 (1941).
[13] H. Roelig and Heidemann, Kunststoffe 7, 125 (1948).
[14] W. Kuhn and O. Künzle, Helv. Chim. Acta 30, 839 (1947); and M. Mooney and R. Gerke, Rubber Chem. Tech. 14, 35 (1941).

curves; however, no systematic study was made of this phenomenon.

The stress-strain properties were measured on an Instron machine† at 25°C and 50 percent relative humidity. For each composition five separate runs were made; the variations in Young's modulus, determined graphically from the initial slope of the stress-strain curve, and the tensile strength are less than 10 percent, but the variations in elongation at break were at times as much as 100 percent for the same composition.

A 5-sec. modulus was measured with a torsional instrument, first described by Clash and Berg.[15] The specimen, which consists of a narrow strip immersed in a liquid bath of constant temperature, is held rigidly at the lower end, whereas the upper end is attached to a rod to which a known torque can be applied. The angular deflection after 5 sec. is noticed and Young's modulus is calculated knowing the torque, the angle of twist, and the dimension of the specimen, using a Poisson ratio of 0.5 to convert from shear to Young's modulus. Shore hardness was measured by the indentation and/or the penetration of a needle-type point under usually variable load into the surface of the plastic composition. The viscosity of the plasticizers was measured with a capillary viscosimeter of the Ostwald-Fenske type. The viscosimeters were immersed in a liquid bath controlled to ±1°C.

The dielectric constant and power factor were measured for all compositions at 1000 c.p.s. as a function of temperature. The measurements were made with a General Radio capacitance bridge using circular specimens, and the surfaces of each specimen was painted with a du Pont silver electrode paint. For a few compositions measurements were also made at 500 and 10,000 c.p.s. The d.c. resistance was measured with a General Radio resistance bridge using General Electric type electrodes. The precision of these measurements was at best ±15 percent, but in many cases it was much poorer because of the high resistance of the samples, which approached frequently the upper limit of sensitivity of the instrument. For all electrical measurements the samples were immersed in an air bath controlled to ±0.5°C.

RESULTS

Figures 2 and 3 show some typical log modulus and dissipation factor data as a function of temperature for various plasticized compositions. Within a certain temperature range the modulus is high, of the order of magnitude of 10^{10} dynes/cm^2, whereas in the higher temperature range the modulus is low, of the order of magnitude of 5×10^7 dynes/cm^2. These two regions, the hard and the rubbery one, are joined by a third one where the modulus changes strongly in a relatively narrow temperature range and this range can be called the transition region. The dissipation factor, or damping,‡ also changes with temperature in a characteristic manner; the damping is low and fairly insensitive to temperature in the hard and the rubbery region, and goes through a definite maximum in the transition region. In the hard and rubbery region the modulus and the damping factor approach a constant value which is fairly independent of plasticizer and plasticizer concentration. It is clear, therefore, that these curves can be characterized quite satisfactorily by two constants, the transition temperature and the half-width.[16] The transition temperature is defined as the temperature where the dissipation factor has its maximum, or where the log modulus-temperature curve goes through an inflection point. The use of transition temperature, as defined above, to characterize the transition region is to be preferred to other temperatures (flex temperature, brittleness, etc.) for the following reasons: (a) it is not an arbitrarily defined quantity, as many of the other temperature values; (b) it is independent of the half-width, and (c) it leads to a simpler relationship with the amount of plasticizer than do the other quantities. The half-width is defined as the temperature difference in degrees centigrade where the dissipation factor has reached half its maximum value. Columns 5

FIG. 1. Torsion pendulum for measurements of dynamic mechanical properties.

† We wish to thank Dr. Hamburger of the Fabrics Research Laboratories for the use of this instrument. A description of this instrument is given by G. S. Burr, Electronics 22, No. 5, 101 (1949).

[15] R. F. Clash and Berg, Ind Eng. Chem. 34, 1218 (1942).

‡ δ = logarithmic decrement = $\pi\frac{G''}{G'}$; $\frac{G''}{G'} = \frac{\text{imaginary modulus}}{\text{real modulus}}$ = dissipation factor.

[16] Dienes and Dexter, Ind. Eng. Chem. 40, 2319 (1948).

and 7 of Table I give the transition temperature and the half-width of all compositions. In a few cases it was not possible to determine the half-width with any degree of reliability, because the absolute value of the maximum could not be determined. In these cases the half-width given is the best estimate possible. Figure 4 shows the dielectric constant and loss factor as a function of temperature for various plasticized compositions. Here again it is possible to characterize these curves by two constants, namely the transition temperature and the half-width. Columns 6 and 8 of Table I show these two quantities for all compositions.

The dependence of the mechanical and electrical transition temperature on plasticizer concentration turns out to be a very simple one, as can be seen from Figs. 5 and 6. For all the plasticizers investigated, with the exception of dilauryl phthalate, which will be discussed later, the transition temperature is a linear function of the volume fraction (c_v) of the plasticizer. For four members of the phthalate homologous series the slope of the $T_t - c_v$ curve is the same, whereas the slope is different for tricresyl phosphate and dibutyl sebacate. The concentration dependence of T_t is independent of frequency, but the absolute value of T_t is dependent on frequency. Finally, it should be pointed out that the transition temperature as a function of the mole fraction of the plasticizer yields a non-linear relationship between these two quantities and also different curves for each plasticizer.

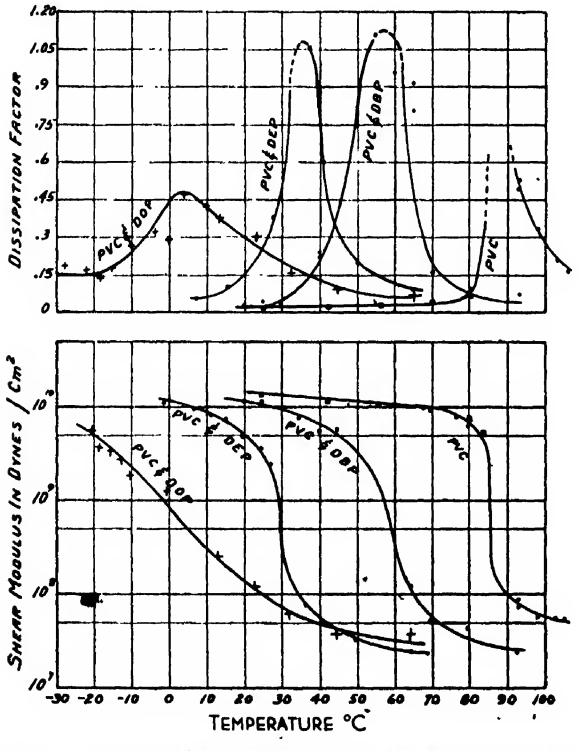


FIG. 2. Dynamic shear modulus and dissipation factor as a function of temperature. 0.158 DBP+PVC; 0.254 DEP+PVC; 0.402 DOP+PVC; PVC without plasticizer (number gives volume fraction).

The dependence of the half-width of the electrical power factor-temperature curve on plasticizer and plasticizer concentration is shown in Fig. 7. The half-width first increases with plasticizer concentration, goes through a maximum at about 0.5 c_v, and then appears to decrease again. For a given plasticizer concentration the half-width increases with increasing chain lengths of the mechanical dissipation factor. It was not possible to determine the relationship between the mechanical half-width and plasticizer concentration with the same accuracy as in the electrical case. However, a correlation between the electrical and mechanical dissipation factors shown in Fig. 8 suggests very strongly that the same conclusions arrived at in the electrical case also hold for the mechanical dissipation factor. It should also be noted that the mechanical dissipation factor is always smaller than the corresponding electrical dissipation factor.

The most interesting feature of stress-strain curves of plasticized polyvinyl chloride compositions (Fig. 9) is the wide difference in the shape of the stress-strain curve as a function of plasticizer and plasticizer concentration. Some stress-strain curves show a definite yield point; in other cases the stress increases rapidly and then remains almost constant over a considerable elongation range before the sample breaks. A more careful analysis of the data shows that the shape of the curves is closely related to the region in which the sample finds itself: in the hard region the slope of the stress-strain curve is very steep and frequently followed by a yield value before break; in the transition region the stress-strain curve slopes strongly, followed by a very flat region with an indefinite yield point, whereas in the rubbery region the stress-strain curve has a small slope which changes little over a very wide range of elongation. The relationship between initial Young's modulus and (a) elongation at break and (b) the tensile strength is shown in Fig. 10. Figure 11 shows the ultimate elongation as a function of the volume fraction of plasticizer; although a slight difference seems to exist for various plasticizers, the number of tests and the precision of the elongation measurement do not justify putting too much emphasis on these differences, and, to a first approximation, one can conclude that the ultimate elongation is very small up to a critical plasticizer concentration at a given temperature and then increases linearly with concentration, irrespective of the type of plasticizer employed. Column 9, Table I gives the ratio of the initial Young's modulus, obtained from the stress-strain curve, to the corresponding dynamic shear modulus obtained with the torsion pendulum, both measured at 25°C.

In studying the creep properties of the various compositions listed in Table I, it was found (a) that repeated creep and recovery runs on the same sample

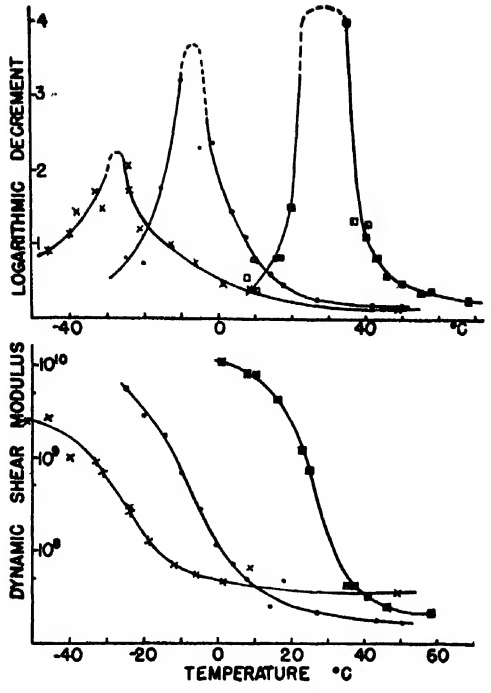


FIG. 3. Dynamic shear modulus and dissipation factor as a function of temperature. ● −0.443 DEP; × 0.458 DBS; ⊡ 0.351 TCP.

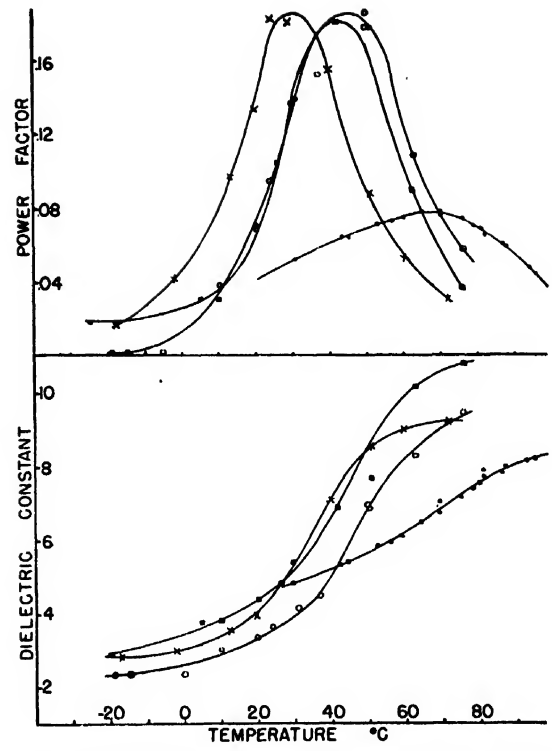


FIG. 4. Dielectric constant and power factor as a function of temperature at 1000 c.p.s. ⊙ .443 TCP; ■ 0.291 DMP; ○ 0.351 TCP; ● 0.368 DLP.

changed the creep properties of the specimen[17] and (b) that the compliance is a function of the applied stress.[4] As a matter of fact, the relationship between load and compliance appears to show significant differences depending on the region in which the material finds itself. In the hard region an increase in the load shifts the compliance-log time curve along the compliance axis toward higher values. In the transition region the slope of the compliance-log time curve decreases with increasing load and in the rubbery range the reverse appears to be true. For these reasons, it seemed more reasonable to study the creep at small and similar elongations by varying the load correspondingly. Some typical results obtained in this manner are shown in Figs. 12 and 13. In the rubbery region the initial elongation is very large, and then changes only slightly with time.§ As one approaches the transition region, the initial elongation becomes less and the change in elongation with time is quite pronounced. In the hard region the change of elongation with time becomes very small, although very few reliable experiments were carried out in this range. In other words, "the curvature" of the elongation-time curve increases as the material proceeds from the rubbery to the transition range and then decreases again going from the transition to the hard region and does not seem to depend very much on the type or amount of plasticizer. Therefore, if one plots the difference in delayed and instantaneous creep as a function of temperature, which was first suggested by Conant and Liska,[18] one obtains a curve very similar to the dissipation factor-temperature curve shown previously and the temperatures of the maxima of both curves are identical. It is believed that the similarity of these two sets of data is more than a coincidence.

The d.c. resistance of the plasticized compositions and of some of the plasticizers was measured as a function of temperature. Over a rather limited temperature range the dependence of the resistance on temperature can be represented by the usual exponential relationships and the slopes of the $\log R - 1/T$ curves which can be represented as an energy of activation, are given in column 10 of Table I. It will be noticed that this quantity increases in all cases with plasticizer concentration, and shows wide variations for different plasticizers at about equal concentration. It should be pointed out, however, that the experimental precision of these measurements was not good enough to establish whether or not the resistance changed in a similar manner as the modulus with changing temperature.

Clash-Berg data and Shore hardness values were

[17] Alfrey, Wiederhorn, Stein, and Tobolsky, J. Colloid Sci. 4, 211 (1949).

§ The initial elongation at a given temperature correlates satisfactorily with the dynamic modulus at the corresponding temperature.

[18] F. S. Conant and J. W. Liska, J. App. Phys. 15, 767 (1944).

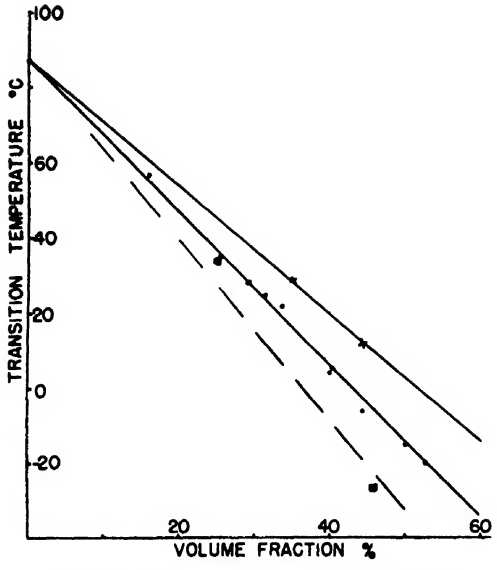


FIG. 5. Dynamic mechanical transition temperature as a function of plasticizer concentration. ● phthalate plasticizers; ⊙ tricresyl phosphate; ⊡ dibutyl sebacate.

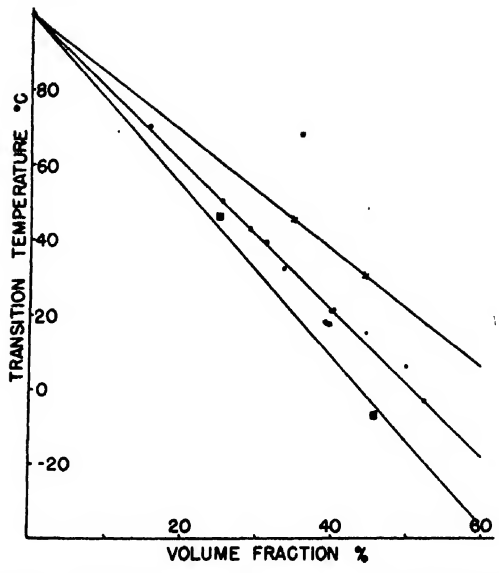


FIG. 6. A.c. electrical transition temperature as a function of plasticizer concentration at 1000 c.p.s. Symbols as in Fig. 5.

obtained as a function of temperature. The shape of the static 5-sec. modulus temperature curve is the same as that for the dynamic modulus temperature curve and the $T_t - c_v$ curves are identical with those shown in Fig. 5, but the absolute values differ by a constant factor. The correlation between Shore hardness and moduli data is not very satisfactory. This is in agreement with previous work on similar compounds[19] and shows the limited usefulness of Shore hardness data to determine moduli of plasticized compositions.

DISCUSSION

During the last few years, several suggestions have been made concerning the mechanism of plasticization; in most cases, these considerations have been concerned with plasticizer efficiency as measured, for example, by the transition temperature, whereas the half-width, or range of plasticization, has not been considered to any great extent. It is the purpose of this section to consider these theoretical attempts in the light of the data presented above.

It has been suggested that the viscosity[5] and the temperature coefficient of the viscosity of plasticizers are of major importance with respect to plasticizer efficiency. However, if the plasticizer efficiency is measured in terms of the transition temperature, then the data given above certainly do not support this contention; see also column 11, Table I. Furthermore, no significant correlation could be established between the temperature coefficient of the viscosity and the temperature dependence of the modulus (or the half-width of the damping curve) in the transition region. Other investigators[6] have tried to show a correlation between the compatibility and plasticizer efficiency; they suggest that the efficiency goes through a maximum as a function of the thermodynamic μ-value. However, it has been shown here that, at least for the first eight members of the homologous phthalate series, the transition temperature is completely independent of the μ-value of the different plasticizers. Furthermore, it is not possible to establish a satisfactory correlation between the μ-value and the slope of the transition temperature-volume fraction curve, not only for the plasticizers studied in this investigation but also for other plasticizer-resin systems reported by other investigators. Nevertheless, the deviation of dilauryl phthalate from the $T_t - c_v$ relationship for all other phthalate plasticizers suggests that the compabitility has some bearing on the transition temperature. Recent investigations[20,21] have brought out rather clearly the relationship between compatibility and plasticizer efficiency: to the extent that the plasticizer is compatible

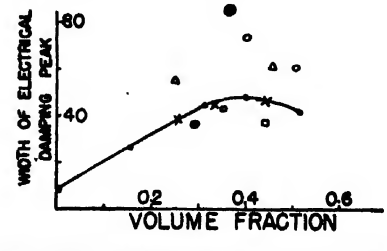


FIG. 7. Half-width of electrical power factor-temperature curve as a function of plasticizer concentration. ■ unplasticized PVC; ⊙ DMP; × DEP; ● DBP; ⊙ DOP; ⊗ DLP; △ DBS; □ TCP; (curve drawn for DBP).

[19] K. H. Reiss, Rubber Chem. Tech. 18, 398 (1945).

[20] W. R. Richards and P. A. S. Smith, J. Chem. Phys. 18, 230 (1950).

[21] Merz, Buchdahl, and Nielsen, High Polymer Forum, ACS Meeting (February, 1949).

with the resin, the dependence of the transition temperature on the volume fraction of the plasticizer is within the same range as the data given above. Beyond, what might be called a critical concentration, the transition temperature then remains essentially unchanged as more of the incompatible plasticizer is added to the resin. In other words, as the plasticizer becomes incompatible with the resin it has only a negligible effect on lowering the transition temperature, that is the plasticizer efficiency approaches zero.|| Neither does there exist a close correlation between the half-widths, at equal volume concentration, and the μ-value; it rather appears, on the basis of the data shown in Fig. 7 that the half-width increases with increasing chain length of plasticizers belonging to a homologous series.

Based on the analogy between the a.c. electrical properties and the dynamic mechanical properties, it has been suggested at various times that the mechanical properties of plasticized polymeric systems can be interpreted on the basis of a simple relaxation theory. Assuming such a mechanism, Alfrey[22] has derived an

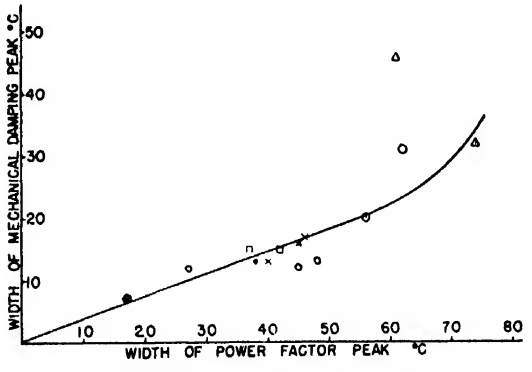


FIG. 8. Correlation between mechanical and electrical half-widths.

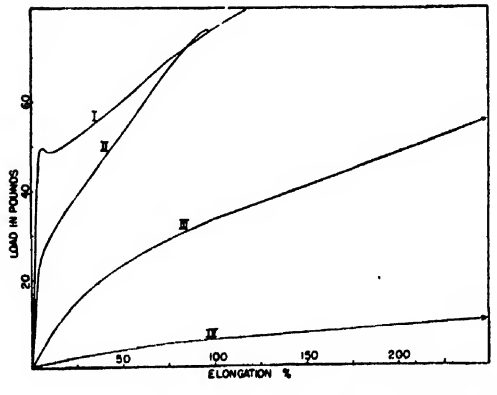


FIG. 9. Typical stress-strain curves of plasticized polyvinyl chloride. I—0.254 DEP; II—0.250 DBS; III—0.402 DOP; IV—0.526 DBP.

|| The μ-value still retains its significance with respect to the compatibility of plasticizers and resins.

[22] T. Alfrey, Jr., J. Chem. Phys. 12, 374 (1944).

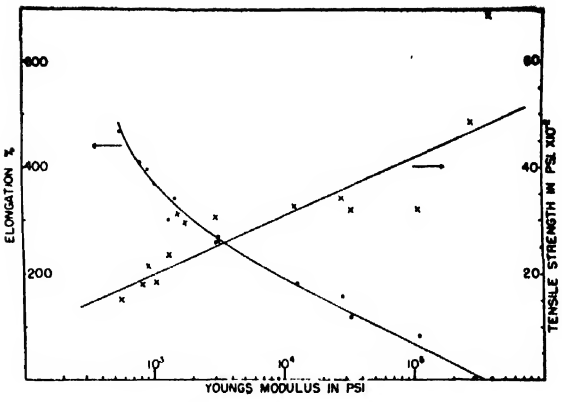


FIG. 10. Ultimate elongation and breaking strength as a function of initial Young's modulus.

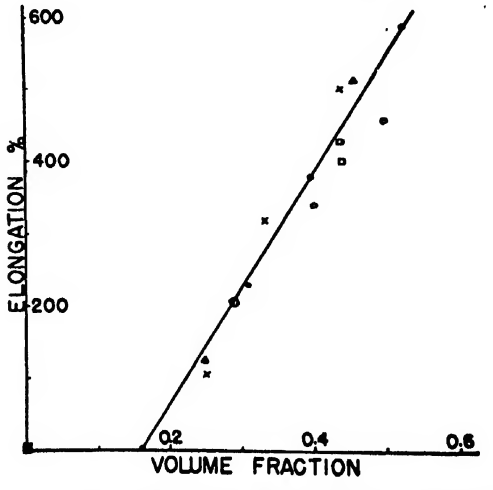


FIG. 11. Ultimate elongation as a function of plasticizer-concentration. ■ unplasticized PVC; ● DBP; ○ DOP; × DEP; ⊙ DMP; □ TCP; △ DBS (curve drawn for DBP).

expression for the compliance as a function of time (or frequency) and molecular weight. If one now considers the transition temperature as an indication of an isoelastic state, one can then derive an expression for the transition temperature as a function of plasticizer concentration, with the result that the reciprocal of the absolute transition temperature must be proportional to the square root of the weight average molecular weight of the mixture. The same relationship has been obtained by Boyer and Spencer,[6] without using Alfrey's compliance equation, by applying Flory's equation of viscous flow of liquid polymers[23] to the transition region in the solid state. A plot of the reciprocal of the absolute transition temperature as a function of the square root of the weight fraction (w_f)** shows that the relationship is more or less linear, although the value for the unplasticized resin does not fall on the linear portion of

[23] P. J. Flory, J. Chem. Phys. 17, 223 (1949); J. Am. Chem. Soc. 62, 1057 (1940).

** W_f is proportional to $\bar{M}w$ over a very wide concentration range for polymers of $\bar{M}w > 100{,}000$.

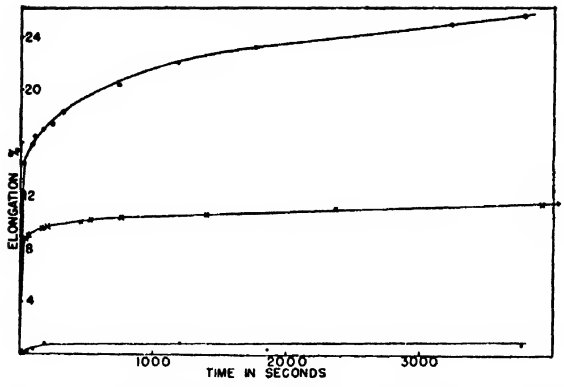


FIG. 12. Elongation as a function of time. Top curve: PVC +0.33 DBP 32°C lower transition range. Middle curve: PVC +0.526 DBP 32°C rubbery range. Bottom curve: PVC+0.158 DBP 22°C hard range.

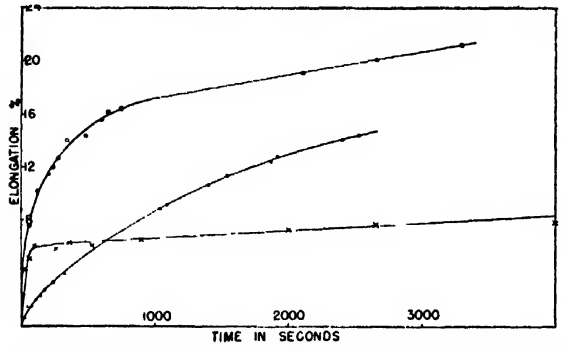


FIG. 13. Elongation as a function of time for PVC+0.254 DEP. Top curve: 32°C lower transition range. Middle curve: 23°C upper transition range. Bottom curve: 45°C rubbery range.

the curve. The square root dependence is due, in both cases, to the use of Flory's equation which, as a result of recent investigations on polystyrene and polyisobutylene, will require some modification.††

Figure 14 shows the transition temperature as a function of the frequency in c.p.s. for several plasticized compositions; the transition temperature changes slowly with frequency, as would be expected on the basis of the relaxation theory. The shape of the curve suggests a very broad relaxation spectrum which has been found to exist for rubber compounds,[24] when investigated over a wide frequency range. However, the frequency dependence of T_t is smaller in the mechanical case than in the electrical.

By and large then, these data could be used to argue in favor of the validity of a relaxation theory were it not for another very important experimental fact, which is the very strong temperature dependence of the modulus or the loss factor in the transition region. It will be recalled that the modulus changes by a factor of 500 to 1000 over a temperature range of 20° to 30°C. This would correspond to an activation energy of the order of 100 kcal. or more, and it is very difficult, if

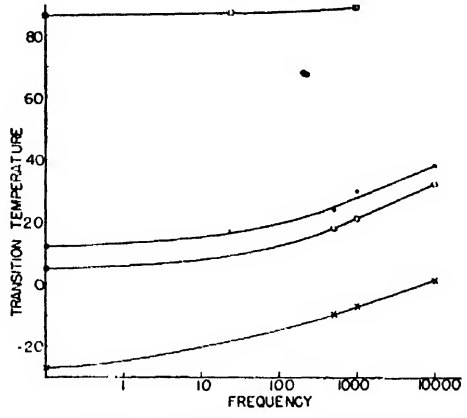


FIG. 14. Transition temperature as a function of frequency. □ unplasticized PVC; ● 0.443 TCP; ○ 0.403 DOP; ⊙ 0.458 DBS.

not altogether impossible, to explain such a high activation energy in terms of a relaxation mechanism. In this connection it seems worth while to point out that the half-width, which is a measure of the temperature dependence in the transition region, is always larger in the electrical case than in the mechanical and on the basis of a relaxation theory one would, at first hand, expect the reverse to be true.

Recently, experimental evidence has been presented in support of the existence of crystalline structure in polyvinyl chloride compositions.[25] One might, therefore, interpret the transition temperature as a melt temperature, and we have tried to fit our data to the equations given by Flory concerning the melting point of polymers.[23] The agreement is not too good using the simplified equation and becomes worse when the more exact equations are being used.‡‡ Furthermore, the existence of a similar transition temperature in completely amorphous polymers, such as polystyrene requires an explanation which is independent of the presence or absence of detectable crystallites. Nevertheless, it seems inescapable to assume some sort of a transition, if for no other reason than to explain the high temperature coefficient of the modulus and loss factor. The nature of this transition is not yet well understood, but it seems to be peculiar to any polymeric or network system.

†† Some years ago, Fuoss (see reference 11) suggested that the product of the mole fraction and a characteristic length of the plasticizer molecule would be a useful quantity to characterize the efficiency of a given plasticizer. When the data given above are treated in this manner, one finds that the transition temperatures of most of the phthalate plasticizers fall on a single curve, but the transition temperature for tricresyl phosphate and dibutyl sebacate fall on a different curve, unless an arbitrary factor is being used to change the length of the plasticizer molecule. This treatment of the data is, therefore, not more advantageous than the treatment given above on the basis of volume fraction.

[24] Witte, Mrowca, and Guth, J. App. Phys. 20, 481, 486 (1949).

[25] Alfrey, Wiederhorn, Stein, and Tobolsky, Ind. Eng. Chem. 41, 701 (1949).

‡‡ The dependence of the transition temperature on the volume fraction instead of mole fraction also argues against the application of a simple melting theory.

Letters to the Editor

X-Ray Tube Producing a Beam of X-Rays Convergent to a Point

C. H. Bachman and S. J. Silverman
Department of Physics, Syracuse University, Syracuse, New York
February 9, 1950

A TUBE has been developed which produces a convergent beam of X-radiation in the form of a hollow cone, the apex of which forms a crossover point in space for the radiation. The radiation density at this crossover point is much greater than at any region in space before or after it.

Figure 1 shows the basic idea of operation of the tube. A filament C is arranged coaxial with target T which is in the shape of a truncated cone. Electrons from the cathode are drawn outward radially to strike the interior face of the cone. X-radiation is emitted in all directions. Much of it is cut off by the shape of the tube. In cross section, that which escapes the tube from the right side is shown as D', that from the left side is D''. Region A is more intense since it has overlapping contributions. No radiation appears at B.

The point F is the region of interest. This point, the apex of the projected conical target, contains like the rest of region A, contributions from all portions of the target area. However, a collimating block may be so placed as to remove all radiation from A and consequently D' and D'' except for a sheath of radiation of controlled thickness. This sheath, a hollow cone of radiation, passes through F, crosses over and diverges beyond.

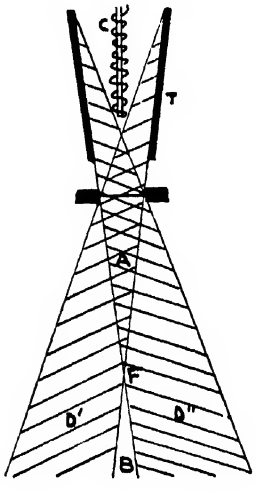


Fig. 1. Schematic ray diagram.

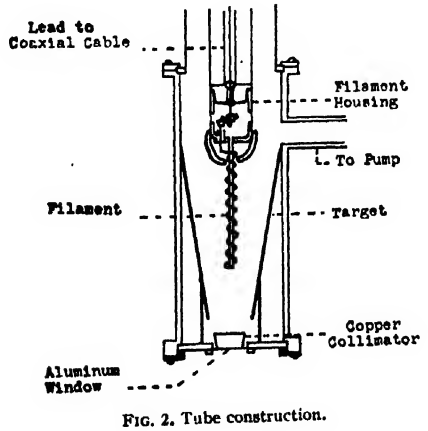


Fig. 2. Tube construction.

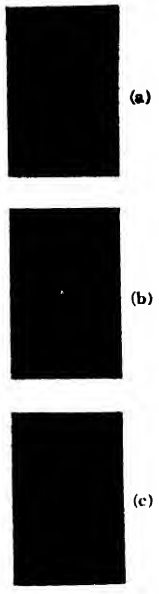


Fig. 3. Beam cross sections. Distance from window: (a) 2.5 cm (before crossover); (b) 7.5 cm (near crossover); (c) 11 cm (after crossover).

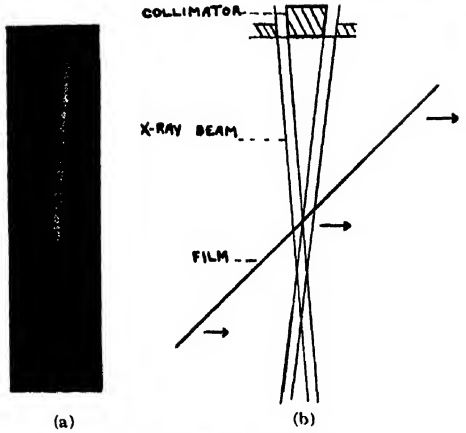


Fig. 4. Crossover presentation.

Figure 2 shows the design features of such a tube and Fig. 3 shows pictures of the beam cross section at several distances from the window. The clover leaf pattern in the scattered outer radiation is due to three radial spacing pins which cross the collimating gap.

Figure 4 was taken by placing a film at 45° to the beam axis and drawing it through the beam. The increased beam density in the crossover region is easily apparent in the negative.

It will be noted that in the tube design no electron focusing problem is encountered; all electrons contribute regardless of where they hit the cone. The very large target area eliminates the customary anode cooling problem or stated conversely, tremendous energy may be used in the tube without being limited by the cooling problem. Also the power rating and x-ray output may be increased by extending the axial length of the conical target since all target areas contribute to the crossover point.

Present efforts are toward explaining some unpredicted radial distribution effects and determining the radiation density distribution in the vicinity of the beam before and after crossover. Other target geometric and methods of utilizing them are also being investigated. More complete results will be published in a future paper.

On Turbulent Jet Mixing in Two-Dimensional Supersonic Flow

D. Bershader and S. I. Pai
Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland
February 16, 1950

At the present time techniques for the measurement of turbulent fluctuations in a supersonic stream are still in the developmental stage and our knowledge of turbulent supersonic flow is correspondingly meager. This note reports a preliminary attempt to investigate the problem by comparing a phenomenological theory of turbulence, extended to the supersonic regime, with some experimental results on the density distribution in such flows. Both the phenomenological type of approach and the measurements in this case deal with the mean properties of the turbulent supersonic flow and lead to some information on the general characteristics of supersonic turbulent flow such as the value of the coefficient of eddy kinematic viscosity and the divergence of jets.

Distributions of mean density in several sections of the mixing region of a rectangular, uniform, fully expanded jet of Mach number 1.7,[1] discharging into the atmosphere, have been calculated from measurements made with the Mach-Zehnder interferometer.* The results thus obtained have been compared with predicted density distributions, calculated by use of a theory developed by the second author in reference 2, which makes use of Reichardt's theory of free turbulence.[2]

In the theory, the coefficient of eddy kinematic viscosity ϵ is assumed to be constant over each cross section of the zone of mixing. Its dependence on X, the distance along the axis of the jet, is found by dimensional analysis to be

$$\epsilon = \epsilon_0 X/L \qquad (1)$$

where ϵ_0 is a proportionality constant having the same dimensions as ϵ, and L is a reference length.

The theory further predicts that the density ρ, in terms of the free stream density ρ_1, i.e., ρ/ρ_1 is a function of y/X, y being the coordinate normal to the flow axis. Hence if we plot ρ/ρ_1 *vs.* $\sigma y/X$, the experimental results for all sections must lie on one curve. σ is an experimentally determined scale factor whose value depends on the divergence or spreading of the mixing region. The larger the value of σ, the smaller the divergence of the jet. Comparison with the theory shows that

$$\sigma = (U_1 L/2\epsilon_0)^{\frac{1}{2}} \qquad (2)$$

where U_1 is the free stream velocity. Other conditions remaining the same, the divergence of the jet decreases with increasing stream velocity.

Figure 1 is a plot of the experimental results for four sections at distances 2, 3, 5, and 7 mm, respectively from the exit of a nozzle approximately 1 by 2 cm in cross section. Superimposed on this is the theoretical curve corresponding to (2). In comparing the experimental data with theory, the following points arise.

(1) Width of the boundary layer at the orifice. The theory assumes zero width of the mixing region at the orifice. In the experiments there exists a 1-mm boundary layer at this point, so extrapolation was employed to obtain the "virtual orifice." This was found to be 10 mm upstream of the actual nozzle exit so that the four test sections were then at X 12, 13, 15, and 17 mm, respectively.

(2) Choice of X axis. Due to the finite dimensions of the nozzle, one would expect a shift of the mixing zone density profile toward the side of the air at rest with increasing X. This is in contrast with a semi-infinite free stream for which a line through the starting point of the jet and parallel to the main flow is an axis of symmetry for the mixing zone and in the present theory corresponds to a density value $\rho/\rho_1 = 0.75$. In the present case, the line $\rho/\rho_1 = 0.75$, chosen as X axis, was found to make an angle θ with the direction of the free stream given by $\theta = 1° 28'$. Gooderum, Wood, and Brevoort,[3] in a similar type of experiment on a jet of Mach number 1.6 found that $\theta = 1\frac{1}{4}°$.

(3) Scale factor. A value of scale factor $\sigma = 17$ was found to give best agreement between theory and experiment. The value of σ in low speed subsonic flow is about 12,[4] showing the smaller divergence of the jet in the present case. (Gooderum *et al.* find $\sigma = 15$.)

From Fig. 1 we see that the experimental points show fairly good agreement with the general shape of the theoretical curve, the deviations becoming more pronounced at the outer edge of the mixing region, which is explained by the fact that in all probability the turbulence is less developed in this region, a phenomenon also found in the case of low subsonic flow.[5]

The calculation of ϵ_0 from the data gives $\epsilon_0 = 7.92$ cm²/sec. To calculate a typical value of ϵ, we find at $X = 17$ mm and $L = 1$ mm that $\epsilon = 134.7$ cm²/sec. which is 1950 times larger than typical values obtained for the laminar coefficient of kinematic viscosity. This is of the same order of magnitude as the ratio of eddy kinematic viscosity to laminar kinematic viscosity for the case of an incompressible fluid.

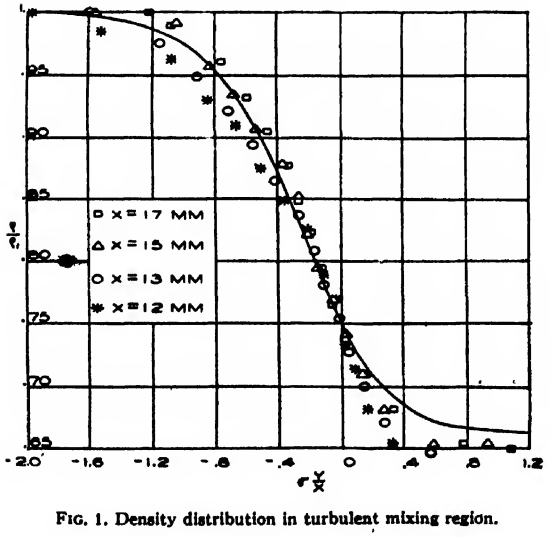


Fig. 1. Density distribution in turbulent mixing region.

[1] D. Bershader, "An interferometric study of supersonic channel flow," Rev. Sci. Inst. 20, 260–275 (April, 1949).
* The experimental data used herein was obtained by the first author in research performed with R. Ladenburg at Princeton University.
[2] S. I. Pai, "Two-dimensional jet mixing of a compressible fluid," J. Aero. Sci. 16, 463–469 (August, 1949).
[3] Gooderum, Wood, and Brevoort, "Investigation with an interferometer of the turbulent mixing of a free supersonic jet," N.A.C.A. Tech. Note 1875 (April, 1949).
[4] H. W. Liepmann and J. Laufer, "Investigations of free turbulent mixing," N.A.C.A. Tech. Note 1257 (August, 1947).
[5] S. Corrsin, "Investigation of flow in an axially symmetrical heated jet of air," N.A.C.A. Wartime Report W-94 (December, 1943).

Use of Brightness Temperatures in Richardson Plots

Henry F. Ivey
Westinghouse Electric Corporation, Bloomfield, New Jersey
February 27, 1950

Quite frequently in experimental investigations it is desired to study the thermionic emission from a material whose spectral emissivity is unknown. In this case the brightness temperature, as determined with an optical pyrometer, cannot be corrected to give true temperature values. It is therefore of interest to investigate the errors in the values of work function, ϕ, and emission constant, A, which result from using brightness

temperatures in making a Richardson plot. The writer is not aware of a published discussion of this kind.

An indication of the magnitude of the effect considered was obtained in the following manner. The emission to be expected from tungsten was calculated at 100° intervals in the temperature range 1200–2800°K, using the constants $\phi=4.52$ volts and $A=60$ amp./cm^2 deg.^{2}. The brightness temperatures corresponding to these points were then determined from the known temperature scale for tungsten.[1] A Richardson plot constructed from these data for emission as a function of brightness temperature showed negligible curvature and yielded values of the constants, as determined by the usual least-squares method, of $\phi'=4.57$ volts and $A'=698$ amp./cm^2 deg.^{2}. It is seen that the apparent work function differs from the true value by only a small amount but the apparent value for the emission constant has been increased by an order of magnitude.

The effect on the emission constants can be predicted theoretically as follows. The brightness temperature, T_B, and true temperature, T, are related by the expression

$$(1/T)-(1/T_B)=(\lambda\log e_\lambda)/c_2=p\log e_\lambda, \tag{1}$$

where e_λ is the spectral emissivity at the wave-length λ used in the pyrometers and natural logarithms are used. The value of the constant p is 4.553×10^{-6} deg.$^{-1}$ for $c_2=1.4385$ cm-deg. and $\lambda=6550$A. When relation (1) is substituted in the Richardson emission equation,

$$J=AT^2e^{-b/T}, \tag{2}$$

the result is

$$J=\frac{AT_B^2e_\lambda^{-bp}}{(1+T_Bp\log e_\lambda)^2}e^{-b/T_B}. \tag{3}$$

It is desired to find the constants b' and A' in the Richardson equation in terms of brightness temperature,

$$J=A'T_B^2e^{-b'/T_B}. \tag{4}$$

The constant b' is given by the relation

$$b'=-\frac{d[\log(J/T_B^2)]}{d(1/T_B)}=T_B^2\frac{d}{dT_B}\log(J/T_B^2), \tag{5}$$

or

$$b'=b-2T_B\left[\frac{T_Bp\log e_\lambda}{1+T_Bp\log e_\lambda}\right]-\frac{T_B^2p}{e_\lambda}\left[b+\frac{2T_B}{1+T_Bp\log e_\lambda}\right]\frac{de_\lambda}{dT_B}. \tag{6}$$

The constant A' can be expressed in terms of b' by elimination between Eqs. (3) and (4).

$$A'=\frac{Ae_\lambda^{-bp}}{(1+T_Bp\log e_\lambda)^2}e^{-(b'-b)/T_B}. \tag{7}$$

In Eqs. (6) and (7), T_B is to be interpreted as the mean of the temperature range over which the measurements are made.[2]

When it is recalled that $e_\lambda<1$, it is seen that the result of the first bracketed term in Eq. (6) is to make the apparent work function ($\phi'=b'/11{,}606$) slightly greater than the true value. The effect of the second term will depend on the sign of de_λ/dT_B and will further increase ϕ' if this quantity is negative (as is the case for tungsten). The effect of the de_λ/dT_B term will be greater the greater the work function. The magnitude of both terms increases for higher values of the measuring temperature and lower values of the emissivity.

In expression (7) for A', the predominant factor is the term

$$e_\lambda^{-bp}=e_\lambda^{-0.528\phi}. \tag{8}$$

It is seen that the apparent emission constant will always be larger than the true value and that the ratio will be larger the higher the work function and the lower the emissivity. From the complete expression (7) it can also be seen that A' increases with increasing values of the measuring temperature.

A check on the numerical example discussed earlier can be obtained by substituting the values of the spectral emissivity (i.e., 0.438) and of de_λ/dT_B (i.e., 2.32×10^{-5} deg.$^{-1}$) for tungsten[1] at 2000°K ($T_B=1860$°K) in Eqs. (6) and (7). The results are $\phi'=4.58$ volts and $A'=746$ amp./cm^2 deg.^{2}, in good agreement with the values $\phi'=4.57$ volts and $A'=698$ amp./cm^2 deg.^{2} obtained numerically. If the de_λ/dT_B term is neglected in the expression for ϕ', one obtains $\phi'=4.54$ volts and $A'=578$ amp./cm^2 deg.^{2}. In general, therefore, the effect of this term is not negligible.

In conclusion it may be said that the use of brightness temperatures instead of true temperatures in the construction of Richardson plots will result in apparent work functions which differ only slightly from the true value, but the apparent emission constant obtained from such a plot may be greater than the real value by more than an order of magnitude. These errors will be greater the higher the measuring temperature and the lower the spectral emissivity of the material in question. In the case of the emission constant the error will also be greater the higher the work function. The variation of the emissivity with temperature also contributes an appreciable effect.

[1] Symposium (sponsored by American Institute of Physics), *Temperature. Its Measurement and Control in Science and Industry* (Reinhold Publishing Corporation, New York, 1941), p. 1318.

[2] In terms of true temperature, Eqs. (6) and (7) may be written

$$b'=b-2T\left[\frac{Tp\log e_\lambda}{1-Tp\log e_\lambda}\right]-(b+2T)\frac{(T^2p/e_\lambda)(de_\lambda/dT)}{1+(T^2p/e_\lambda)(de_\lambda/dT)}, \tag{6'}$$

$$A'=A(1-Tp\log e_\lambda)^2e_\lambda^{(b'-2b)p}e^{-(b'-b)/T}. \tag{7'}$$

Stroboscopic Mapping of Time-Variable Fields

L. Marton and D. L. Reverdin
National Bureau of Standards, Washington, D. C.
March 8, 1950

SEVERAL publications appeared[1–3] on the mapping of steady electrostatic or magnetic fields using the deflections of a viewing electron-beam. We have contemplated for some time the possibility of extending the method to include time-varying fields by the use of a stroboscopic technique. A need for this modification arose in an extensive series of studies of the static field of a cut-off magnetron.[4] In these experiments an electron beam was sent axially through the static magnetron in order to view the space-charge field. The filament heating current, however, generated a magnetic field which masked the space-charge field completely. To eliminate the effect of the filament heating current, the heating

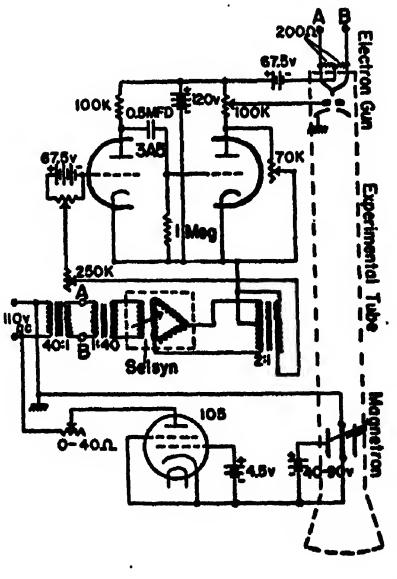


FIG. 1.

was produced by a half-wave rectifier, and the observation of the space-charge distribution was restricted to the other half-cycle.

Figure 1 shows a schematic circuit diagram. Pulse operation of the observing electron beam was achieved by a square-wave modulation of the electron gun bias potential.

An interesting observation while carrying out these experiments was that the image appeared considerably sharper when the stroboscopic method was used than in a steady-state observation. No complete explanation for the sharpening of the image can be offered at this time, but there is good indication that the pulsing of the observing beam reduced the effect of 60-cycle stray fields always present in the laboratory.

The stroboscopic method is applicable in principle to the observation of any repetitive, time-variable fields up to very high frequencies. At very high frequencies there are two serious limiting factors. One is the speed with which the current of the observing beam can be modulated; the second is due to the transit time of the electrons in the field to be observed. The first limitation is the less serious one, since by beam deflection methods instead of biasing of a control electrode, very high frequency pulses can be generated. The second limitation is the more serious one because it requires an increasingly stiff electron beam to keep the transit time low while the sensitivity of the method varies inversely with beam stiffness. A vague estimate of the practical upper frequency limit at which the method can be applied without undue experimental difficulties can be set at about 300 Mc, the present upper limit of oscillograph operation.

[1] L. Marton, J. App. Phys. 19, 687 (1948).
[2] L. Marton, J. App. Phys. 19, 863 (1948).
[3] L. Marton and S. H. Lachenbruch, J. App. Phys. 20, 1171 (1949).
[4] Doctoral dissertation submitted in February, 1950 by D. L. Reverdin to the George Washington University, Washington, D. C. Part of it is scheduled to be published in an early issue of this Journal.

Announcement

From time to time, the Institute has received inquiries from subscribers for information as to where and how they may have their copies bound. After investigation the Institute has found that this may be best accomplished by selecting a capable bindery that will work to our specifications and produce a well-bound volume at as low a price as possible. The necessary arrangements have been completed with the Book Shop Bindery, 308 West Randolph Street, Chicago 6, Illinois.

The volumes will be bound in the best grade of washable buckram with facsimile gold stamping on the spine and the subscriber's name in gold on the front cover. A distinctive color for each of the Journals has been selected as listed below. Advertising matter and covers are removed before binding. The price of $3.00 per volume applies to any of these journals regardless of size. Issues, comprising complete volumes, should be sent by prepaid express or parcel post to the Book Shop Bindery. Bound volumes will be shipped approximately 30 days after receipt. Return transportation will be prepaid when remittance accompanies order.

No deviation in color or specification can be allowed in view of the favorable price.

Physical Review	Green
Reviews of Modern Physics	Orange
Journal of the Optical Society	Brown
Journal of the Acoustical Society	Orange
American Journal of Physics	Black
Review of Scientific Instruments	Grey
Journal of Chemical Physics	Blue
Journal of Applied Physics	Red
Physics Today	Purple

Journal
of
Applied Physics

Volume 21, Number 7 July, 1950

Results of Recent Hypersonic and Unsteady Flow Research at the Langley Aeronautical Laboratory

JOHN V. BECKER

Chief of Compressibility Research Division, Langley Aeronautical Laboratory, National Advisory Committee for Aeronautics, Langley Air Force Base, Virginia

INTRODUCTION

THE extremely high speeds and altitudes now contemplated in supersonic aircraft design involve many new compressible flow problems which, only a few years ago, were either unrecognized or of purely academic interest. This situation has naturally resulted in a greatly increased concentration of effort in basic compressible flow research at establishments such as the NACA. The purpose of this paper is to review briefly some recent results obtained at Langley in projects of this character dealing with hypersonic and unsteady flows. Similar exploratory work in small-scale facilities is in progress at the Ames and Lewis Laboratories of the NACA.

THE 11-INCH HYPERSONIC TUNNEL

The eventual need for a wind tunnel capable of speeds in the range of 5 to 10 times the speed of sound was envisioned in 1945, and we undertook the development of a pilot model of such a facility at that time. The pressure ratios and other important design factors for hypersonic tunnels were then unknown, and this uncertainty pointed towards a blowdown type of tunnel sufficiently flexible in design to permit easy modification. A schematic drawing of the facility is shown in Fig. 1. The combination of a 50-atmosphere supply tank and a vacuum receiving tank affords a starting pressure ratio of several thousand. It was recognized that large reductions in the operating pressure ratio might be achieved through the use of a variable second throat provided that adverse flow separation effects did not occur. Provision was therefore made in the design for quickly reducing the second throat area after starting of the flow.

Two nozzles were designed by the characteristics method for $M=7.0$; one was a conventional single-step type as indicated on Fig. 1, and the second was a two-step design which expanded first in the vertical plane to a Mach number of about 4.5 and then in the horizontal plane to $M=7.0$. The test sections in each case are 10.55 in. deep by 10 in. wide. The single-step nozzle was found to produce acceptably uniform flow at the test section and a mean Mach number of 6.9 as determined from static and total-pressure measurements in the stream. Difficulty was encountered in the two-step nozzle caused by non-uniform boundary layers at the entrance to the second step.

During the preliminary design studies, it became obvious that the temperature in the test section would drop to the liquefying temperature of air at a Mach number of about 4.5 if the air in the settling chamber were maintained at room temperature. For higher speeds, the air temperature would fall far below the liquefying temperature, even if large degrees of supersaturation were assumed. Experience with water-vapor condensation suggested that considerable supersaturation might exist in air. However, the uncertainty as to what would actually happen in air led us to incorporate a heater which can maintain stagnation air temperature up to 1300°F abs. The possibility of air liquefaction can thus be eliminated in our tunnel for Mach numbers up to approximately 8.0. Incorporation of the heater has the obvious additional advantage of making possible studies of temperature effects not only in connection with the condensation phenomena, but also in heat transfer research in this tunnel. As an incidental advantage, the heater also appreciably lengthens the test runs. A water-cooled heat exchanger installed ahead

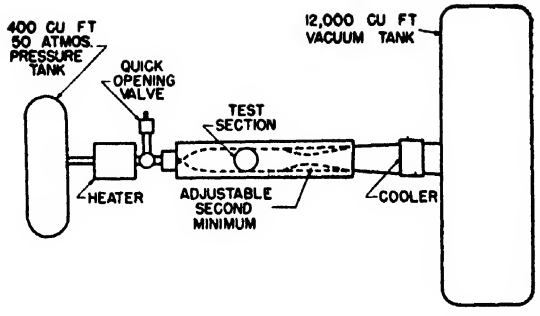


FIG. 1. Schematic drawing of NACA 11-in. hypersonic tunnel.

of the vacuum tank has a similar beneficial effect on the running time. Runs of about 25 sec. duration are obtained without the use of the adjustable second throat. With the second throat immediately adjusted to a reduced area after starting, runs of 60 to 90 sec. are obtained, the time varying with model support configuration.

The tunnel functioned satisfactorily during its first runs on November 26, 1947.

AIR CONDENSATION EXPERIMENTS

Pressure-distribution data have been obtained along the parallel sidewalls of the nozzles of the 11-in. hypersonic tunnel for stagnation temperatures ranging from high values for which condensation is impossible to low values for which the components of air must either be highly supersaturated or condensed. It was found that even for the lowest stagnation temperatures the longitudinal pressure distributions were smooth; no condensation shock was apparent. In this respect, our data agree with results obtained in the 5-in. hypersonic tunnel at California Institute of Technology and in the 2.5-in. $M=4.5$ tunnel at Princeton. Had it been impossible for us to vary temperature, the lack of any apparent condensation shock probably would have led us to the conclusion that no condensation was occuring. Further analysis of the pressure data, however, reveals that, while no discontinuities occurred along the nozzles, there was a significant change in the pressure at all stations along the nozzle when the temperature was allowed to fall below saturation values. Figure 2 shows this effect for various stations along the two nozzles. The left diagram of Fig. 2 shows the results for three stations in the forward part of the nozzle at which the Mach numbers at high temperatures were 5.47, 5.75, and 6.09. The right diagram shows the results for the test-section stations of the two nozzles investigated where the Mach numbers as determined from wall static pressures were 6.56 and 6.76. The pressure ratio at any of these stations was not much affected by decreasing the temperature until the saturation region was reached. Here a marked increase in pressure (decrease in Mach number) was noted. In the entire first step of the two-step nozzle in which the temperature never fell below saturation values, there were no appreciable pressure variations with temperature. These results seem to indicate that condensation, if it is actually present, is a gradual process occurring along the entire region of the nozzle where the temperatures are appreciably below the saturation value. The possibility that these pressure ratio changes are due to viscous or heat transfer effects rather than condensation has not yet been eliminated. However, tests made over a range of stagnation pressures at temperatures above the saturation level indicated no appreciable Reynolds number effects. Instead of the indicated decrease in Mach number with decreasing T_0 in the saturation region (Fig. 2), the effect of the corresponding Reynolds number increase produced a small increase in Mach number.

In recent years, the nucleation theory developed by Volmer,[1] which successfully predicts the condensation of water vapor, has been used by several investigators,[2–4] for example, to estimate the degree of supersaturation of oxygen that might be expected. These calculations, when applied to our conditions, predict high degrees of supersaturation terminated by a condensation shock. An assumption common to all of the theoretical work is that no foreign particles are present which might act as condensation nuclei and thus reduce the probability of high degree supersaturation. In our air stream, there are three possibilities: carbon dioxide crystals, water crystals, and oxide dust from the heater. Water is present only to the extent of about two parts in a million, and thus probably can be neglected. It is also doubtful whether enough oxide particles are present

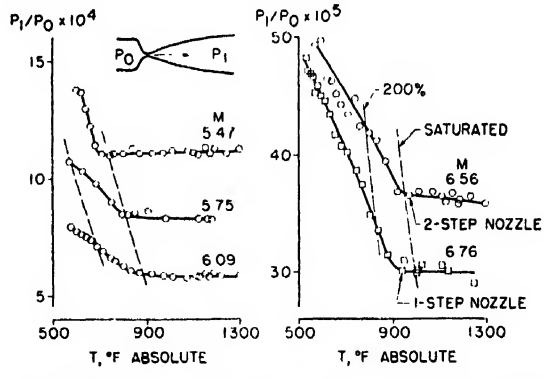


FIG. 2. Effect of stagnation temperature on the static pressure at each of several stations along the nozzle. (The stations are designated in terms of the roughly constant Mach number existing at each station for temperatures above saturation.)

[1] M. Volmer, *Kinetik der Phasenbildung* (Theodor Steinkopff Verlag, Dresden, 1939).

[2] A. E. Puckett and R. Schamburg, "Hypersonic wind-tunnel progress," Report No. 5 (U. S. Army Contract W-04-200 Ord-1963), Guggenheim Aero. Lab., California Institute of Technology, August 8, 1946.

[3] P. Wegener, "On the experimental investigation of hypersonic flow," U. S. Naval Ordnance Lab. Memo. No. 9629, July 29, 1948.

[4] S. M. Bogdonoff and L. Lees, "Study of the condensation of the components of air in supersonic wind tunnels. Part I—Absence of condensation and tentative explanation," Princeton University Aero. Engineering Laboratory Report No. 146, May 25, 1949.

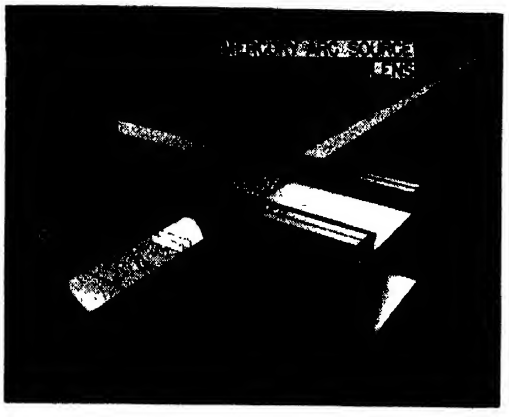


FIG. 3. Set-up of light-scattering experiment.

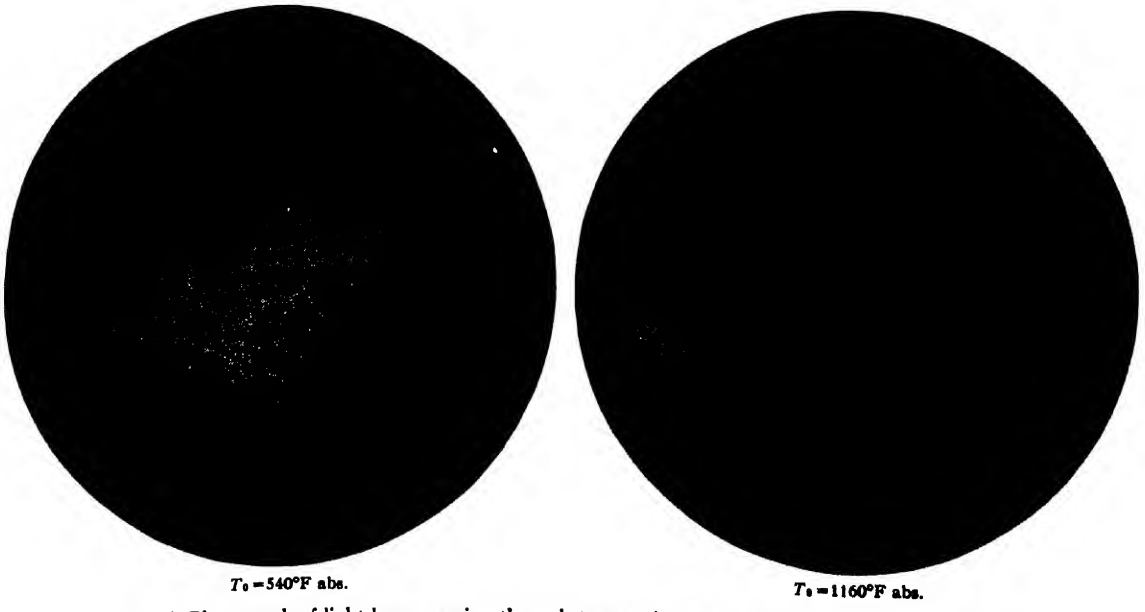
to be significant. If the carbon dioxide solidifies without becoming highly supersaturated, a significant number of solid particles would form in the flow at temperatures somewhat below those corresponding to the saturation of oxygen. Under these circumstances, the theoretical predictions of high supersaturation and condensation shock would have to be modified.

A simple optical experiment has been made which is of interest in connection with the condensation question. Figure 3 shows the experimental setup. An approximately parallel beam of light is passed through the test-section windows with the test chamber darkened. For the no-flow condition, all that can be seen are the reflections where the beam passes through the windows. When the stagnation temperature is at its usual elevated value, 1160°F abs., there is no evidence of any particles in the flow, the oxide particles apparently not being numerous enough to show up (Fig. 4). When the flow is not heated, however, strong scattering of the light beam occurs suggesting the presence of a large number of small liquid or solid particles. Near the walls in the high temperature region of the boundary layer, the light scattering disappears entirely. It is conceivable, of course, that this phenomenon is caused by CO_2 particles rather than by liquid oxygen or nitrogen particles. The possibility has also been advanced that the scattering is caused by liquid oxygen particles of subcritical size, as considered in the Volmer theory. Another conceivable explanation is that photo-nucleation is occurring. This test, therefore, does not provide completely conclusive evidence of condensation of the components of air.

Additional experimental work on the condensation problem is planned but there is a limit to what can be accomplished in the tunnel. It appears to us that cloud chamber techniques can be used to advantage in the study of this problem. Another possibility suggested by Donaldson at Langley is to utilize the sudden expansion in a suitably designed shock tube to simulate the rapid changes in temperature and pressure of a supersonic nozzle. The experimental information obviously needed includes degrees of supersaturation, condensation time factors, and the effects of gas mixtures and foreign nuclei on the condensation phenomena.

FLOW VISUALIZATION AT LOW PRESSURES

Our experience with the 11-in. tunnel indicates that the limit of usefulness of the schlieren method of flow visualization is reached at a pressure of about 1 mm of mercury for slender test models at $M=7.0$. To obtain significant results at these conditions, we use a sensitive system in which the light is passed through the test section twice. A single 10-in. mirror is used located close to one side of the tunnel and ground to the rela-



$T_0 = 540°F$ abs. $T_0 = 1160°F$ abs.

FIG. 4. Photograph of light beam passing through test section taken from same viewpoint as Fig. 3.

tively long radius of curvature of 20 feet. In terms of Mach number, the estimated limiting value for this schlieren system is about $M=10$ for a 50-atmosphere stagnation pressure. For higher hypersonic tunnel speeds, and also for low density aerodynamic research at all speeds, other methods of flow visualization are needed.

We were fortunate to have Dr. Joseph Kaplan as a consultant on this problem during the summer of 1948. He suggested use of the little-understood phenomenon of the afterglow that persists in nitrogen which has been excited by an electrical discharge as a means of flow visualization. Acting on this suggestion, Williams and Benson set up a low pressure supersonic nozzle in which the afterglow principle was applied successfully for the first time[5] in August 1948. The apparatus used is shown on Fig. 5. Nitrogen is ducted through the exciting discharge tube, through a grounded screen which greatly reduces the number of ions present but affects the afterglow only slightly, and thence into the supersonic nozzle. The brightness of the glow can be increased by increasing the energy of the exciter, either by the addition of more arcs or by increasing the power. It is interesting to note that the grounded screen has little effect on the brightness. Figure 6 shows the nozzle flow for three pressure ratios less than that required for starting, which was about 20 to 1. The uniform brightness of the afterglow after starting and the patterns about a 15° double-wedge model are illustrated in Fig. 7. The high pressure region behind the shock has greater brightness than the main stream. Separation occurs at the point of maximum thickness of the wedge-shaped model at the low static pressure of this test, 3 mm of mercury.

The nitrogen afterglow appears to be usable at static pressures as low as 100 microns and perhaps lower. We have also successfully used other gases including dry air, oxygen, and argon, although their afterglows are not as bright as that of nitrogen. A small amount of iodine vapor appreciably increased the brightness of nitrogen afterglow. It is probable that other combinations can be found having more intense afterglows.

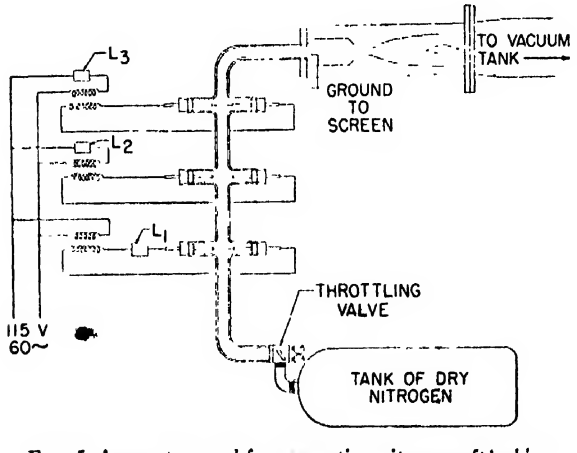


FIG. 5. Apparatus used for generating nitrogen afterglow in a supersonic nozzle.

[5] T. W. Williams and J. M. Benson, Preliminary Investigation of the Use of Afterglow for Visualizing Low-Density Compressible Flows, NACA RM I9A24a, February 15, 1949.

The question arises as to whether the presence of the glow has any aerodynamic effects on the flow. Studies made by Rayleigh[6] indicated that the decay of the glow was accompanied by a high rate of heat release, a phenomenon which would be expected to produce important aerodynamic effects in a supersonic nozzle. This possibility has recently been investigated by Benson at Langley. He finds that in our application for which the glow has a life of about one sec., the temperature rise is of the order of 30° to 90°C. This corresponds to a temperature rise of only a small fraction of a degree during the transit of the gas through a supersonic nozzle. Thus the aerodynamic effect is expected to be negligible. Benson also finds no appreciable change in viscosity due to the presence of the glow. Schlieren and afterglow photographs obtained for static pressures of 6 to 9 mm of mercury correlate satisfactorily. In summary, the afterglow technique at its present stage of development appears to be a reliable and convenient method of flow visualization for low pressure aerodynamic research.

BOUNDARY-LAYER EFFECTS AT HYPERSONIC SPEEDS

It is well known that the effect of the boundary layer on the pressure distribution about aerodynamic bodies at both subsonic and supersonic speeds is restricted to the rear or trailing portion of the body. In this region, adverse pressure gradients occur which cause the boundary layer to thicken greatly or to separate from the surface, with associated large effects on the surface pressures. Over the forward part of the body, the boundary layer for Reynolds numbers in the practical range is too thin to have a measurable effect on the pressures. This fact is well established for the entire speed range for which pertinent data are available in aeronautical literature.

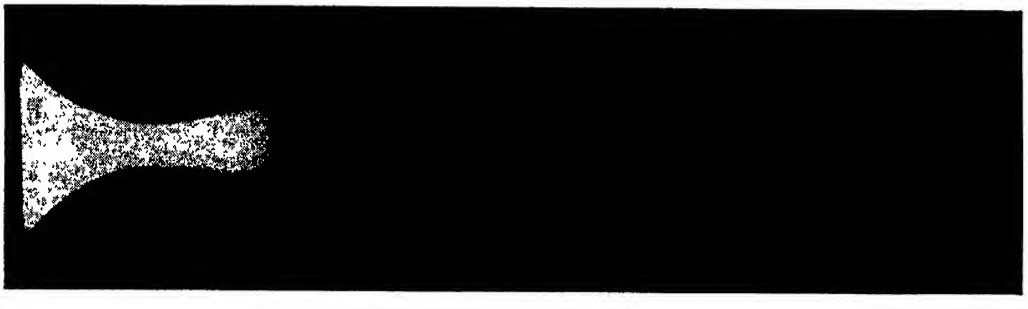
The change from low supersonic to hypersonic speeds involves a many-fold increase in boundary-layer thickness for a given value of the Reynolds number. Recent calculations and tests in the 11-in. hypersonic tunnel at $M=6.9$ have brought to light the fact that the thickness becomes so great as to cause a large distortion of the pressure diagram over the forward portion of airfoils. The usual separation effects over the rearward portion are also present.

Figure 8 shows the calculated growth of the laminar boundary layer on a flat plate as a function of Mach number. The calculations are based on the usual compressible-flow boundary-layer assumptions with the added assumption of a linear velocity profile. The latter

[6] S. K. Mitra, *Active Nitrogen—A New Theory* (Indian Press, Ltd., Calcutta, 1945).

(A) Pressure ratio approximately 9 to 1

(B) Pressure ratio approximately 10 to 1



(C) Pressure ratio approximately 11 to 1

FIG. 6. Nozzle flow patterns obtained with nitrogen afterglow at pressure ratios below that required for starting.

assumption is justified on the grounds that over the forward part of a flat plate at our test Reynolds number the boundary layer is known to be laminar and that theoretical considerations[7,8] reveal that the profile is expected to be nearly linear at hypersonic speeds. (See upper diagram on Fig. 8.) Furthermore, the calculated chordwise growth of the displacement thickness is insensitive to the profile shape chosen. Figure 8 shows that the displacement thickness at $M=7$ is about 10 times that at $M=1.0$ for a constant value of $x/(R_x)^{1/2}$, where x is the distance from the leading edge of the plate and R_x is the Reynolds number based on the length x. The increase is associated with the large temperature rise in the boundary layer at high Mach numbers and to the corresponding viscosity increase near the surface.

FIG. 7. Flow pattern for wedge-shaped test model obtained with nitrogen afterglow. $M=2.6$. Absolute static pressure, 3 mm of mercury.

[7] A. Busemann, "Gas Stromung mit Laminarem Grenzschlict Entlang Einer Platte," ZAMM 15, S23, 1935.

[8] Th. von Karman and H. S. Tsien, "Boundary layer in compressible fluids," J. Aer. Sci. 5 No. 6 (April, 1938).

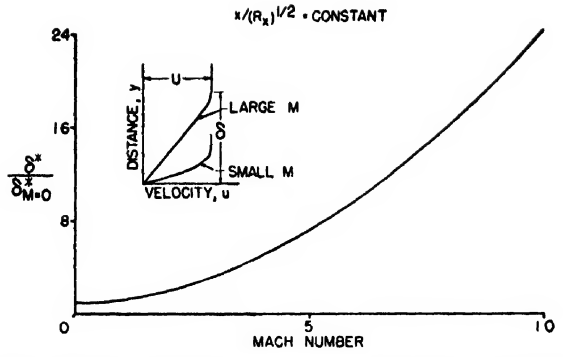


FIG. 8. Effect of Mach number on the displacement thickness of the laminar boundary layer on a flat plate.

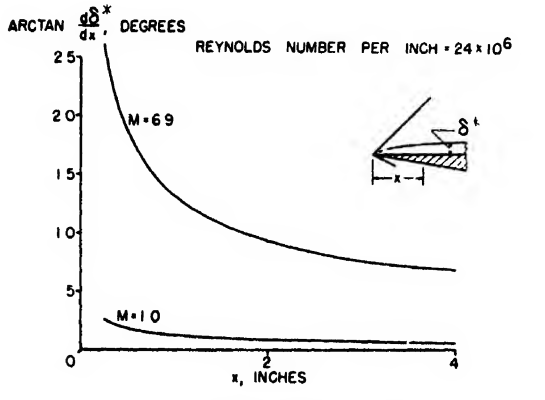


FIG. 9. Slope of the effective surface of the boundary layer along a flat plate.

Figure 9 presents the effective change in slope of the surface of the plate due to growth of the boundary layer. The effect, of course, is a maximum near the leading edge. At the 0.20-in. station, for example, the slope increase is about 2.3° or the equivalent of a four-percent-thick wedge.

The calculated pressure increments corresponding to the slopes of Fig. 9 are plotted on Fig. 10 as fractions of the stream static pressure. For comparison, the results of actual measurements for a flat surface parallel to the stream in the 11-in. hypersonic tunnel are also given. It will be seen that the pressures near the leading edge are markedly altered by the boundary-layer effect; a pressure rise of about 0 7 of stream pressure occurs at the 0.25-in. station. The agreement between theory and experiment appears surprisingly close when it is considered that no account was taken of the effect on the boundary-layer development of either the bow shock or the streamwise pressure gradient.

A schlieren photograph of the flow over a flat surface, taken at a static pressure of 5 mm of mercury, abs., is shown on Fig. 11. The significant part of this photograph is the upper surface of the wedge, alined so as to be accurately parallel to the flow. The lower shock in Fig. 11 is due to the inclined lower surface and has no significance in the present discussion. The shock caused by the boundary layer can be seen to be curved near the nose as required by the effectively curved surface corresponding to the chordwise growth of the boundary layer. The thick boundary layer itself can also be seen; the apparent thickness agrees approximately with the calculated values.

Two obvious practical effects of this phenomenon are to increase the pressure drag of hypersonic airfoils and to alter appreciably the characteristics of leading-edge control devices.

UNSTEADY FLOWS

Research in the field of unsteady flows is receiving increasing attention partly because of the possibilities of using unsteady flows to advantage, for example, in compressors, pulse-jet engines, and high speed wind tunnels, and partly because of serious adverse effects due to unsteady flows on the performance of supersonic inlets, diffusers, and other aircraft components. In the field of useful applications is the familiar "comprex" compressor, designed by Seippell, in which moving compression and expansion waves perform the func-

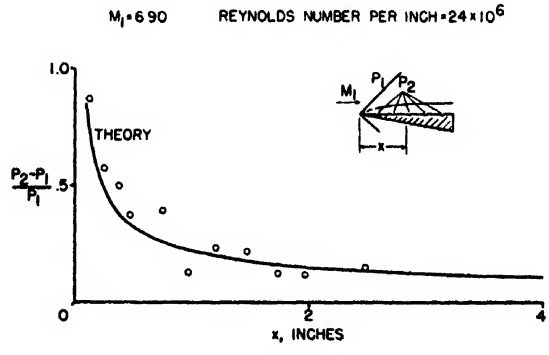


FIG. 10. Effect of boundary layer growth on the pressure distribution along a flat plate.

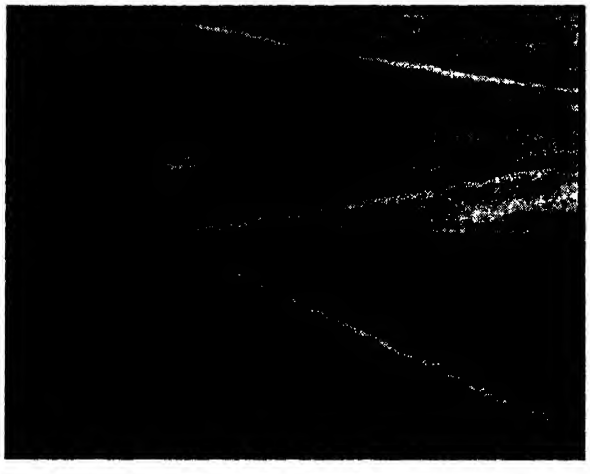


FIG. 11. Schlieren photograph of the boundary layer and shock on a flat surface aligned with the flow. $M = 6.9$. Absolute static pressure, 5 mm of mercury.

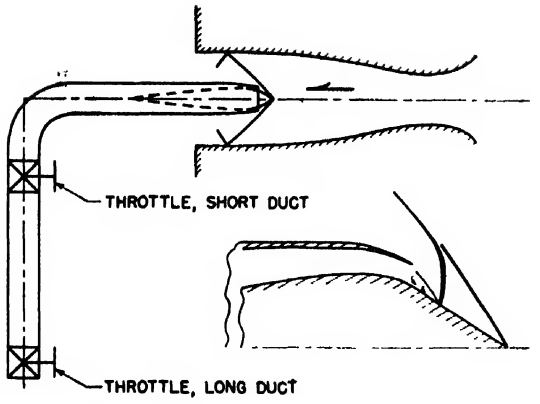


FIG. 12. Supersonic inlet test configuration and sketch of shock-induced separation on conical nose.

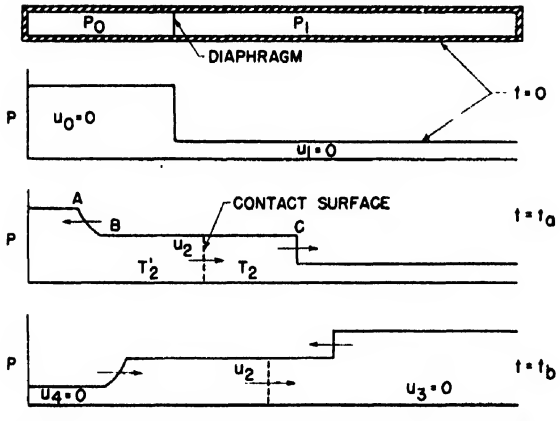


FIG. 13. Illustration of shock tube pressure distributions.

tions of a mechanical piston.[9] Another useful unsteady flow device is the shock tube, which has been used, for example, by Smith at Princeton to study shock reflections[10] and by Geiger and others at the University of Michigan as an intermittent flow wind tunnel.[11]

Adverse unsteady flow phenomena on aircraft wing sections arising from shock boundary-layer interaction at high speeds have been studied for a decade or more and these phenomena continue to present some of the most difficult problems of high speed research. Recently Ferri at Langley has discovered an interesting case of unsteady flow due to shock boundary-layer interaction at the entrance of an axially symmetric supersonic inlet. Figure 12 illustrates the experimental set-up used. The inlet is shown inserted in the test region of a supersonic nozzle which produced a Mach number of 1.90. Back pressure in the duct was varied by means of a throttle, and tests were made with two alternate throttle locations as shown on Fig. 12. The flow pulsations occurred when it was attempted to reduce the mass flow into the inlet by closing the throttle and thereby forcing the normal shock to move from its design location (just inside the inlet) to an external position on the conical central body, thus permitting spillage of some of the approaching air. The external position of the shock is illustrated on the lower sketch of Fig. 12. The basic cause of the pulsations was the abrupt occurrence of shock-induced separation resulting from failure of the boundary layer on the cone to negotiate the strong normal shock in its external position. This separation produced an abrupt reduction in the mass flow entering the inlet and a corresponding drop in back pressure in the duct. The sudden drop in back pressure is equivalent in its effect to a sudden opening of the throttle and thus the shock was momentarily drawn into the inlet again, after which the process repeated itself. The phenomenon was initiated solely by the external shock-induced separation and was in no way connected with internal diffuser or burner disturbances. As would be expected, however, the length of the internal duct had a marked effect on the frequency and character of the pulsations once they had started. The "long duct" configuration (Fig. 12) had lower frequency but larger amplitude pulsations. For both duct configurations the pulsations were of such violence as to prohibit any useful applications of this inlet. Various means of preventing periodic separation on the cone are being studied and there is promise that this type of unsteady flow can be eliminated.

As a basis for experimental study of some of the more difficult unsteady flow problems previously mentioned, Huber has investigated in some detail the phenomena in a simple shock tube.[12] The objectives of this work were to develop suitable pressure-measuring equipment

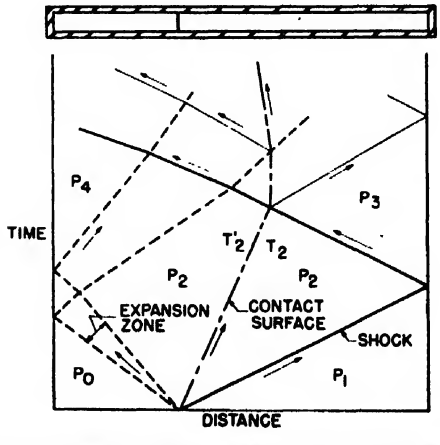


FIG. 14. Position histories of shock-tube phenomena.

[9] C. Seippell, Pressure Exchanger, U. S. Patent Office 2,399,394, Pat. April 30, 1946.

[10] L. G. Smith, Photographic Investigation of the Reflection of Plane Shocks in Air, OSRD Report No. 6271, November 1945.

[11] F. Geiger and C. W. Mantz, "The shock tube as an instrument for the investigation of transonic and supersonic flow patterns" Contract N6-ONR-232, Project M720-4, Engineering Research Institute, University of Michigan, June, 1949.

[12] Huber, Fitton, and Delpino, "Experimental investigation of moving pressure disturbances and shock waves and correlation with one-dimensional unsteady flow theory." NACA TN 1903, July, 1949.

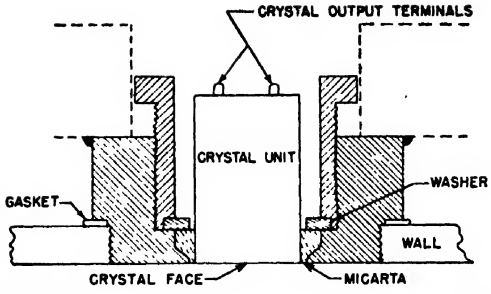


FIG. 15. Method of mounting pressure measuring crystal in wall of shock tube.

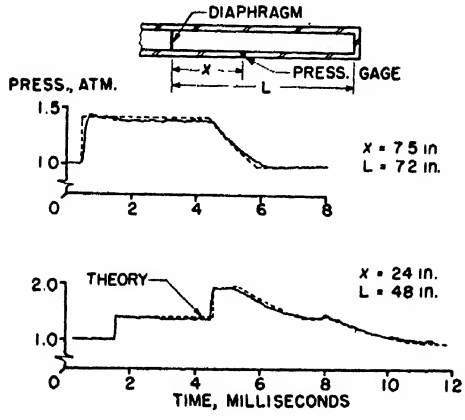


FIG. 16. Pressure histories in the shock tube for two crystal positions and two lengths of tube.

and to determine how closely the one-dimensional theory predicts the actual phenomena in the tube. For the benefit of those unfamiliar with the shock tube, Fig. 13 has been prepared to illustrate its action. The tube has two chambers separated by a diaphragm and containing a gas (or two gases) at different pressures. When the diaphragm bursts, a shock proceeds towards the low pressure end of the tube at a low supersonic speed and a rarefaction wave proceeds in the opposite direction at the speed of sound. Between these disturbances, the pressure is constant at the value P_2 and a flow is produced of velocity u_2 in the direction of the low pressure end. The surface of contact between the gas originally at high pressure and that at low pressure represents a temperature and entropy discontinuity, since the air to the right has increased in temperature due to the shock passage, while that at the left has decreased in temperature due to the action of the rarefaction. When the shock strikes the end of the tube, it is reflected and proceeds in the opposite direction, reducing the flow velocity in the tube to zero and increasing the pressure. The rarefaction wave has a similar action at the opposite end of the tube. It is important to note that any compressive disturbance behind the shock quickly catches up with it because of the velocity u_2 and the higher speed of sound behind the shock. The initial rarefaction wave tends to spread out, however, because the velocity of advance of a wave

FIG. 17. Schlieren photographs of the shock-tube phenomena at various times (seconds) after bursting of the diaphragm.

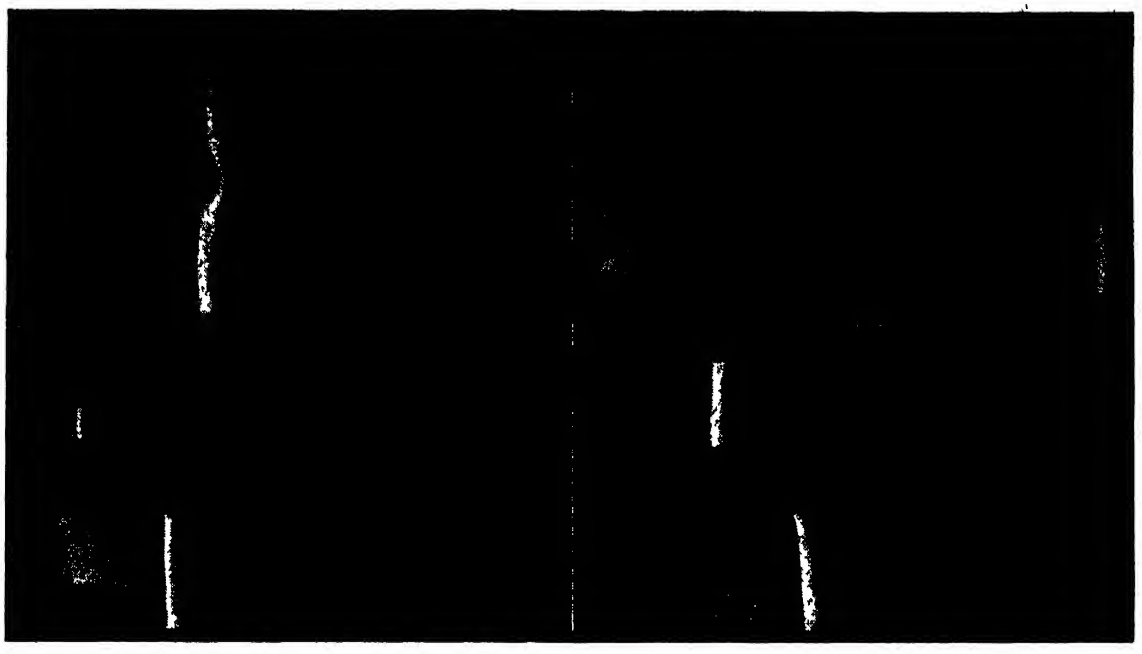


FIG 18 Schlieren photographs of a shock traveling from left to right against subsonic flow in a channel containing an asymmetrical divider.

behind the rarefaction front is retarded by u_2 and by the reduced speed of sound.

Figure 14 shows the theoretical positions of the disturbances as functions of time. Interactions among the various disturbances occur and after the first cycle the flow becomes increasingly complex. The attenuation of the disturbances due to wall friction has been studied by Donaldson[13] and found to be very small in a tube of these proportions for any one cycle. The method of characteristics provides a tool for calculating these interactions and reflections. Neglecting friction and following the work of Guderley and others, Hess[14] has successfully applied the characteristics method to several much more difficult cases—for example, to the starting of flow in a variable cross-section supersonic nozzle when a diaphragm or quick-acting valve is opened at the throat.

Huber's tests were made with air, using an initial pressure ratio of 2.0. The internal cross section of the shock tube was 3×3 inches in size. The tube length varied from 56 to 104 in. the high pressure chamber being 32 in. long and the low pressure chamber being variable in length from 24 to 72 in. Figure 15 illustrates the pressure measuring crystal used. This is a commercial unit $\frac{5}{8}$ in. in outside diameter selected for its properties of high acoustic impedance, high sensitivity, and absence of cavity resonance. The crystal did not perform well when mounted in rubber, when a small air cavity and leakage were present, or when it was slightly misaligned.

The pressure traces produced by the crystal when mounted in each of two positions on the low pressure side of the diaphragm are presented in Fig. 16. In the upper diagram, the crystal was near the diaphragm and the trace shows the pressure rise due to the shock to take place over a measurable time increment. This result is caused primarily by failure of the diaphragm to burst instantaneously as a perfectly plane surface. In this connection, it is of interest to mention that of several materials tried for the diaphragm, the best was tracing paper—both because of its disintegration into many small pieces and because it had less stretch than the other materials. The number of thicknesses used was selected to match the pressure differential. Figure 16 also shows the reflected rarefaction zone and the slight rise in pressure due to reflection of the rarefaction from the temperature discontinuity. In general, the data are in reasonable agreement with the theoretical predictions.

The lower diagram of Fig. 16 was taken with the pick-up farther away from the diaphragm and with a shorter chamber. Here the shock is more nearly plane as would be expected at a greater distance from the initial burst because lagging disturbances have had time to catch up with the shock front. The small discontinuity in the theoretical curve at $t = 5.2$ and 8,1 milliseconds is due to reflection of the shock from the temperature discontinuity at the contact surface. Again there is reasonable agreement with the theory.

[13] C. duP. Donaldson and R. D. Sullivan, "The effect of wall friction on the strength of shock waves in tubes and hydraulic jumps in channels," NACA TN 1942, September, 1949.

[14] R. V. Hess, "Study of unsteady flow disturbances of large and small amplitudes moving through supersonic or subsonic steady flows," NACA TN 1878, May, 1949.

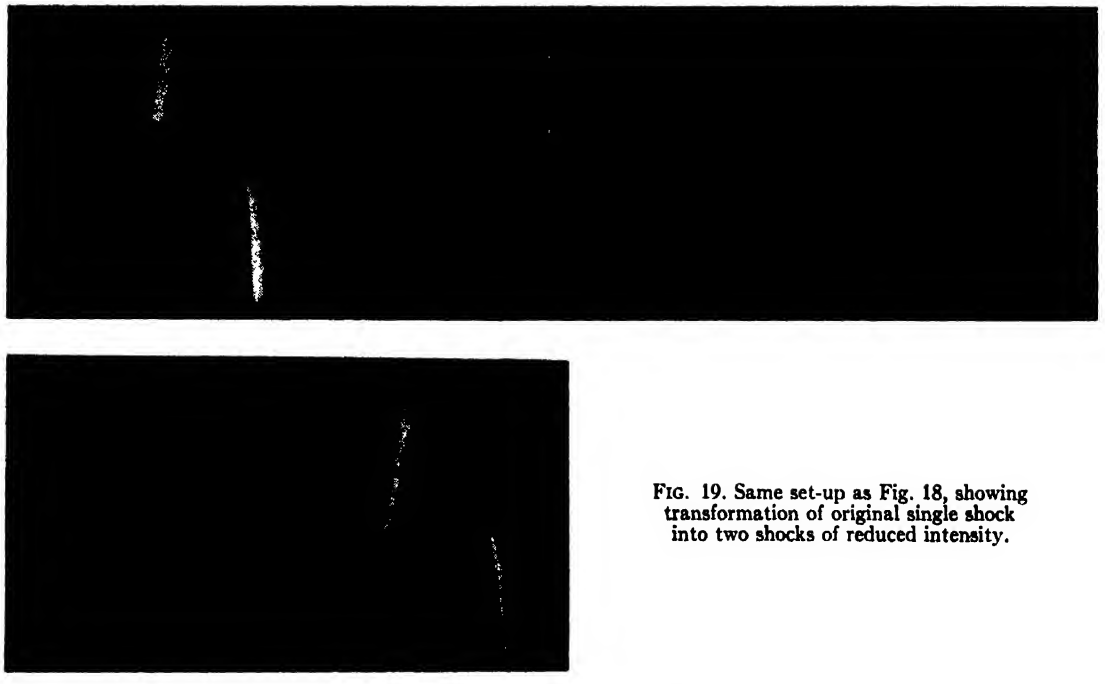


FIG. 19. Same set-up as Fig. 18, showing transformation of original single shock into two shocks of reduced intensity.

Schlieren photographs of the disturbances are shown on Fig. 17. The first picture, ($t=0.4\times10^{-3}$ sec.), taken just downstream of the ruptured diaphragm, shows the "shock" as a wide band. This explains why it did not appear as a discontinuity in the pressure trace. A few tube widths farther downstream, ($t=1.1\times10^{-3}$ sec.) the shock appears as a nearly plane discontinuity, although all of the reflections generated earlier have not yet caught up with it. The third picture, ($t=2.0\times10^{-3}$ sec.), shows the shock immediately after reflection from the end of the low pressure chamber. The oblique disturbances are believed to be reflections of the shock from the walls which were generated as a result of initial curvature of the reflected shock produced by deformation of the 11/16-in. thick end plate. The fourth picture ($t=2.7\times10^{-3}$ sec.) shows the reflected shock passing through the temperature discontinuity. This "discontinuity" is actually a turbulent band—in this case about one tube width in thickness. The finite thickness and turbulence are due mainly to irregularities introduced by the bursting of the diaphragm. Distortions of the shock in passage through this region are evident.

It is concluded from Huber's work that the only significant departures from the one-dimensional shock tube theory arise from the finite time and distortions associated with the rupturing of the diaphragm.

The final subject to be discussed briefly is related to the unsteady flow problem of supersonic diffusers. It is well known that the normal shock in supersonic diffusers oscillates in position due to pressure disturbances from the subsonic part of the diffuser itself or from the combustion chamber. The distance between the mean position of the shock and the throat of the diffuser must be increased as the amplitude of the oscillations increase. For maximum diffuser pressure recovery, the shock must be located as closely as possible to the diffuser throat and this requires therefore that the oscillations be minimized or, in the ideal case, eliminated.

A scheme proposed by Ferri for reducing the strength of the disturbances which generate the oscillations of the shock has been the subject of a preliminary investigation, results of which are illustrated in Figs. 18 and 19. These figures show schlieren photographs of a compressive disturbance (weak shock) traveling to the right against a subsonic flow ($M\approx0.2$) in a tube of square cross section. In the middle of the tube is an asymmetrically shaped two-dimensional body which splits the subsonic flow into two smaller channels of unequal area and velocity. The disturbance divides into two parts upon contacting this central body (upper right picture of Fig. 18). When the two shocks emerge into the channel at the forward end of the divider (lower pictures of Fig. 18) the upper shock lags behind the lower due to the greater retardation provided by the higher speed subsonic flow in the upper passage. Upon leaving the divider (upper pictures of Fig. 19) each shock tends to spread across the entire channel, forming a final configuration of two disturbances (lower picture of Fig. 19) each having about half of the intensity of the original shock. Of course, the rear shock will eventually overtake and coalesce with the leading shock if the duct is long enough. Numerous variations of this scheme suggest themselves. For example, a greater number of channels can be used to reduce further the strength of the initial disturbance or to produce any desired pulsation frequency. Considerable further theoretical and experimental research is required to explore the many possible applications of this principle.

Contributed Original Research

Modulation Conversion in a Wave Guide

PAUL S. ROGELL
Hazeltine Electronics Corporation, Little Neck, New York
(Received November 29, 1949)

Theory and design equations are presented for the conversion of a microwave, narrow-band f-m signal into very nearly a pure a-m signal within an ordinary wave guide of specified dimensions. Frequency relationships and guide dimensions are arranged so that the f-m side bands are shifted with respect to the carrier until they become a-m side bands at the guide output. A complex intelligence signal may be transmitted with negligible distortion if it is used to amplitude modulate an intermediate carrier which frequency modulates the microwave carrier.

IT is often desirable to utilize an amplitude-modulated signal at microwave frequencies. Since the generation of a microwave frequency-modulated signal is not difficult, the desired operation could be carried out if one possessed the means of converting a microwave f-m signal into an a-m signal. Such a conversion can be accomplished in an ordinary wave guide.

The basic principle employed is the fact that in a wave guide operated near cut-off the phase velocity of a single-frequency signal varies widely with the frequency of the signal. Down such a guide, the low frequency components of a signal travel faster than do the high frequency components. It is possible to adjust the position of the cut-off and the length of the guide so that the frequency components of an f-m wave at the input of the guide will rearrange themselves to form very nearly a pure a-m wave at the output.

The vectorial representation of a narrow band f-m wave, consisting of only a carrier and the first two side bands, is shown by Fig. 1a. It can be seen from the figure that as the carrier rotates, the two side bands add up so as to impart an alternating component to the speed of rotation of the carrier. In the figure, ω_c is the carrier angular frequency and ω_m is the angular frequency of a single-frequency modulating signal.

Suppose that the f-m signal of Fig. 1a be introduced into a wave guide operated near cut-off. Each of the three components of the signal will be changed in amplitude and phase upon arriving at the guide output. Let us arrange the parameters of the system so that the lower side band is shifted 450° and the upper side band 270°, each with respect to the carrier. Expressions for the relations that must hold among the parameters to achieve these phase shifts are developed in Appendix A. Also, let us assume that the side bands are equally attenuated. Actually, the resultant side bands are not quite equal, and the output contains a little phase modulation (see Appendix B).

Under the above conditions for phase and amplitude changes, the output signal may be represented by the vector diagram of Fig. 1b. The output is a pure a-m signal, since the two side bands add up so as to impart an alternating component to only the amplitude of the carrier.

It is generally desirable to use a complex wave, not a single frequency, as the transmitted information. To accomplish such a transmission, it would seem natural to merely allow ω_m to be variable. However, the phase shifts of the side bands with respect to the f-m carrier are intimately associated with the frequency separation of those side bands from the carrier. Varying ω_m, therefore, would destroy the proper phase relations necessary to produce an a-m signal. The difficulty can be overcome by using the signal of frequency ω_m as an intermediate carrier that is amplitude modulated by the intelligence. In this case, the output signal is a microwave carrier amplitude modulated by an a-m intermediate carrier. The intelligence introduces a distortion term, but as long as the highest intelligence frequency is less than one-tenth the intermediate-carrier frequency, less than one percent distortion is introduced.

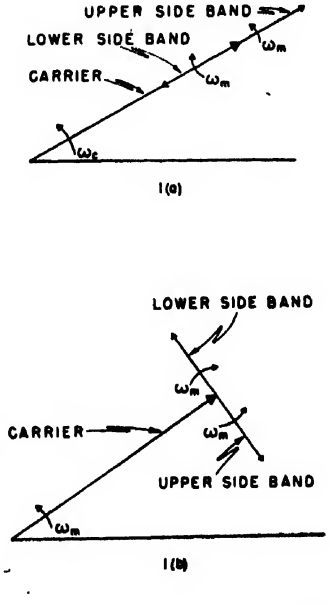


FIG. 1. Idealized vector diagrams for the f-m signal into the guide, and the a-m signal at the guide output.

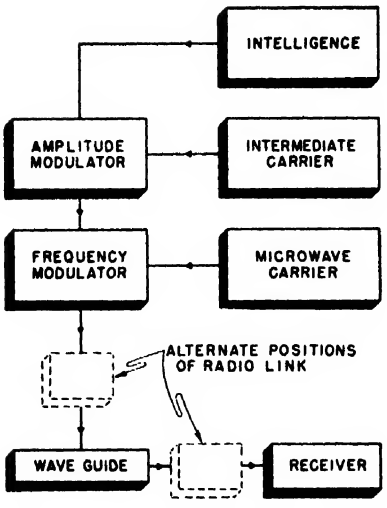


FIG. 2. Block diagram of the complete system.

The case of a variable intelligence frequency is analyzed in Appendix C.

The receiver used in the system could consist merely of two stages of crystal detection together with a suitable amount of amplification.

A simplified block diagram of the entire system is shown in Fig. 2. If the radio link were placed after the wave guide, then an a-m carrier would be transmitted, and a simple receiver could be employed. If the radio link were placed before the wave guide, then an f-m carrier would be transmitted, and the wave guide coupled to the crystals would act as a frequency-modulation detector.

Design equations for the wave-guide dimensions and for frequency relationships are developed in the Appendices. Either the guide output or the input should be variable in longitudinal position so as to secure exactly the desired length. As another design consideration, it would probably be necessary to use crystal stability when generating the intermediate and microwave carriers.

Suppose an amplitude-modulated signal is desired at a carrier frequency of 3000 mc. Let the wave guide be rectangular, carrying the TE_{10} mode. Let the guide cut-off be 2960 mc, necessitating a transverse guide dimension of 5.06 cm. Then the design equations show that the intermediate-carrier frequency must be 33.2 mc, and the guide length must be 1.30 meters. The band width required would be 73.0 mc. Intelligence frequencies ranging from 0 to 3.3 mc would yield less than one percent amplitude distortion in the guide output. There would be a maximum frequency deviation at the guide output of 0.58 mc. This deviation is not large enough to affect the crystal detection in the receiver. The amplitude of the f-m side bands in the guide output relative to the a-m side bands would be 0.0091.

Under the above conditions, the guide output is nearly pure amplitude modulation, since the amplitude of the f-m side bands is less than one percent of the a-m side band amplitude, and there is less than one percent quadrature-component amplitude distortion.

APPENDIX A

Let us investigate the relations that must exist among the system parameters if the side bands of the f-m carrier are to be shifted in phase by the proper amounts. The phase constant of a single-frequency wave in a guide may be expressed in terms of the frequency of the wave and the cut-off frequency of the guide.[1] We may write,

$$\beta_1=(\omega_c/c)\{[1+(\omega_m/\omega_c)]^2-(\omega_{co}/\omega_c)^2\}^{\frac{1}{2}}, \tag{1}$$

where β_1 is the phase constant of the upper side band, ω_{co} is the guide cut-off frequency, and c is the speed of light. MKS units are used throughout.

Let δ be the difference in frequency between the lower side band and guide cut-off. That is,

$$\delta=\omega_c-\omega_{co}-\omega_m.$$

Neglecting terms of the order of $(\omega_m/\omega_c)^2$ compared to one, we may write,

$$(\omega_{co}/\omega_c)^2=[1-(\omega_m+\delta)/\omega_c]^2\approx 1-2[(\omega_m+\delta)/\omega_c],$$

and

$$[1+(\omega_m/\omega_c)]^2\approx 1+2\omega_m/\omega_c.$$

Substituting the approximate forms just derived into Eq. (1)

$$\beta_1\approx 1/c[2\omega_c(2\omega_m+\delta)]^{\frac{1}{2}}.$$

Similarly, it can be shown that

$$\beta_2\approx 1/c(2\omega_c\delta)^{\frac{1}{2}},$$
$$\beta_c\approx 1/c[2\omega_c(\omega_m+\delta)]^{\frac{1}{2}},$$

where β_2 and β_c are the phase constants of the lower side band and the carrier respectively.

The ratio of the phase shift in the lower side band to the shift in the upper side band is

$$\frac{(\beta_c-\beta_2)l}{(\beta_1-\beta_c)l}\approx\frac{(\omega_m+\delta)^{\frac{1}{2}}-\delta^{\frac{1}{2}}}{(2\omega_m+\delta)^{\frac{1}{2}}-(\omega_m+\delta)^{\frac{1}{2}}}, \tag{2}$$

where l is the length of the guide. This ratio varies from one to about 2.42 as ω_m varies from zero to $(\omega_c-\omega_{co})$. $5\pi/2$ and $3\pi/2$ were chosen as the smallest phase shifts that would accomplish the desired modulation conversion and still have a ratio between one and 2.42. Setting the right side of Eq. (2) equal to 5/3, and solving for ω_m, we find

$$\omega_m=(240/49)\delta,$$

or

$$\omega_m=240/289(\omega_c-\omega_{co}). \tag{3}$$

Substituting the values for β_c, β_2, and ω_m into the equation for the phase shift of the lower side band,

$$l=5\pi/2(\beta_c-\beta_2),$$

we find

$$l=\frac{17\pi c}{4[2\omega_c(\omega_c-\omega_{co})]^{\frac{1}{2}}}. \tag{4}$$

The guide length and intermediate carrier frequency are determined from Eqs. (3) and (4) given the carrier and cut-off frequencies.

If TE_{10} propagation is used in a rectangular guide, then the transverse guide dimension, a, is[2]

$$a=\pi c/\omega_{co}.$$

---

[1] See, for example, S. Ramo and J. R. Whinnery, *Fields and Waves in Modern Radio* (John Wiley and Sons, Inc., New York, 1944), p. 297.

[2] See reference 1, Table 9.04.

APPENDIX B

This section considers the attenuation of the carrier and the first two side bands between guide input and output. The attenuation constant, α, for a wave of angular frequency ω, in a square guide propagating the TE_{10} mode is[2]

$$\alpha=\frac{\omega_{co}}{\pi c}\left(\frac{\omega\epsilon}{2\sigma}\right)^{\frac{1}{2}}\frac{[1+2(\omega_{co}/\omega)^2]}{[1-(\omega_{co}/\omega)^2]^{\frac{1}{2}}},$$

where ϵ is the dielectric constant within the guide, and σ is the conductivity of the guide walls.

The ratios of the output to input amplitudes for the carrier, the upper side band, and the lower side band will be $\exp(-\alpha_1 l)$, $\exp(-\alpha_2 l)$, and $\exp(-\alpha_3 l)$, respectively, where α_1, α_2, and α_3 are the attenuation constants for waves of frequencies ω_c, $\omega_c+\omega_m$, and $\omega_c-\omega_m$, respectively. The output a-m wave may be written,

$$e_{out}=AE_c\{\sin\omega_c t-(Bk/2)[\cos(\omega_c+\omega_m)t-D\cos(\omega_c-\omega_m)t]\}, \quad (5)$$

where E_c is the input carrier amplitude, k is the degree of modulation, and A, B and D are attenuation factors, that is,

$$A=\exp(-\alpha_1 l),$$
$$B=\exp[-(\alpha_2-\alpha_1)l],$$
$$D=\exp[-(\alpha_3-\alpha_2)l].$$

The constant A would merely change the apparent carrier amplitude and the constant B would change the apparent degree of modulation. Neither A nor B would influence the purity of the a-m output signal.

The constant D is the relative amplitude of the side bands. Let us consider the two unequal side bands to be two equal side bands of amplitude D plus a single side band of amplitude $1-D$. The maximum phase deviation of the output carrier would be $(1-D)/2$ radians, so the maximum frequency deviation would be $(1-D/2)\omega_m$. The amplitude of the first-order f-m side bands is proportional to $J_1[(1-D)(\omega_m/2)/\omega_m]$ or to $J_1[(1-D)/2]$, where J_1 is the Bessel function of first kind, first order. For small arguments,

$$J_1[(1-D)/2]\approx(1-D)/4$$

and the amplitude of the f-m side bands relative to the a-m side bands is $(1-D)/4D$. Using the values of the numerical example in the text, together with

$$\sigma_{copper}=5.80\times10^7$$
$$\epsilon=8.85\times10^{-12}$$

it is found that

$$D=0.965$$

The relative strength of the f-m side bands is 0.0091, and the maximum frequency deviation is 0.58 mc.

APPENDIX C

Let ω_s be the variable frequency of an intelligence signal, and let this signal amplitude-modulate the intermediate carrier. The new amplitude of the intermediate carrier, E_m', is

$$E_m'=E_m(1+M_s\cos\omega_s t)$$

where E_m is the original intermediate-carrier amplitude, and M_s is the degree of modulation by the intelligence. The instantaneous frequency of the f-m input, ω_i, is now

$$\omega_i=\omega_c[1+M_m E_m(1+M_s\cos\omega_s t)\cos\omega_m t],$$

where M_m is the degree of modulation by the intermediate carrier. The guide input, e, is

$$e=E_c\sin\left(\int_0^t\omega_i dt\right).$$

Using standard forms to integrate, and neglecting terms of order of magnitude ω_s^2 compared to those of order of magnitude ω_m^2, we have

$$e=E_c\sin\left\{\omega_c t+\frac{M_m E_m\omega_c}{\omega_m}\left[(1+M_s\cos\omega_s t)\sin\omega_m t-\frac{M_s\omega_s}{\omega_m}\sin\omega_s t\cos\omega_m t\right]\right\}. \quad (6)$$

The last term in Eq. (6) is a distortion term. Its amplitude relative to the amplitude of the desired signal is ω_s/ω_m, and it is in quadrature with the desired signal. Therefore, if ω_s/ω_m never exceeds one-tenth, the distortion will be less than one percent.

The band width is

$$BW=2(\omega_m+\omega_{s_{max}})\approx2.2\omega_m,$$

where $\omega_{s_{max}}$ is the highest intelligence frequency. Substituting from Eq. (3),

$$BW=1.83(\omega_c-\omega_{co}). \quad (7)$$

Equation (7) is a design equation. Given the carrier frequency, cut-off may be chosen from considerations of band width, and the tolerable amount of f-m signal in the output (see Appendix B).

Measurement of Self-Diffusion of Silver without Radioactive Tracers*

G. C. KUCZYNSKI
Sylvania Electric Products, Inc., Bayside, New York
(Received September 23, 1949)

A new method dispensing with the use of radioactive tracers has been developed for measurement of volume self-diffusion. The method is based upon the relationship of the rate of sintering of metallic wires to the coefficients of self-diffusion. The values of the self-diffusion coefficients for silver as measured by this method are in good agreement with those obtained by the radioactive tracers method.

SELF-DIFFUSION implies the diffusion of the atoms of a given substance into itself. Until recently, the detection of such a process was possible only by measurement of the rate of penetration of the radioactive isotope into another stable isotope of the same chemical element.

This method when first introduced by von Hevesy[1-3] was restricted to the study of diffusion of the naturally radioactive elements such as lead[4] and bismuth.[5] The discovery of artificial radioactivity removed this obstacle and allowed the determination of self-diffusion coefficients of copper, gold, silver and zinc using improved methods.[6,7]

In the previous work of this author,[8] the relationship of the radius x of the interface between the sintered particles to the coefficient D of volume self-diffusion was established.

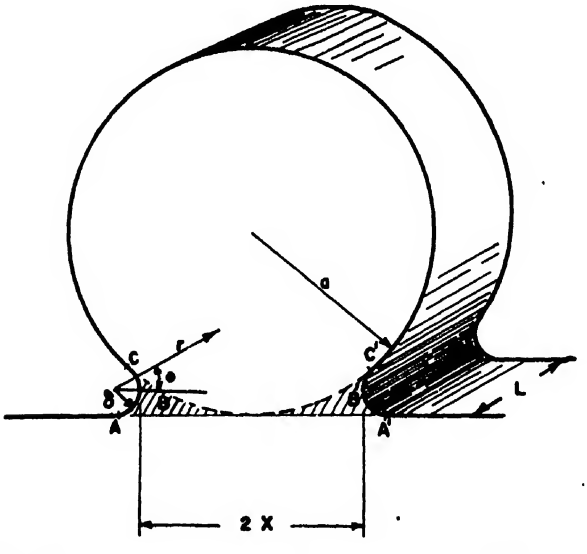


FIG. 1. Schematic representation of a wire bonded to a cylinder.

This relationship is as follows:*

$$D=k/(40\gamma\delta^3)(T/t)(x^5/a^2), \qquad (1)$$

where k is Boltzmann's constant, γ the surface energy of the metal δ the interatomic distance, T the temperature of sintering on the Kelvin scale, t time of sintering and a the original radius of the particle. Work with small metallic particles is laborious and does not yield accurate results. The examination of the graphs contained in the above mentioned paper reveals a definite linear relationship between $(x^5/a^2)T/t$ and $1/T$. However, the spread of points is too large for accurate determination of the coefficient of diffusion. In his communication to the American Physical Society,[9] the author was able to report some of the diffusion coefficients of silver determined by a new method consisting of sintering of silver wires to cylindrical silver pieces. The detailed description of this method and results obtained is the subject of this paper. The new method allows to determine the self-diffusion of most metals without use of radioactive tracers.

THEORETICAL CONSIDERATIONS

The wire in contact with the plate of the same metal forms a system which is not in equilibrium because its total surface free energy is not a minimum. During heating, the space between the wire and the plate

TABLE I. Sintering of 0.005-inch silver wires to silver cylinders.

T°C	t hrs.	$x\times10^{-4}$ cm	$a\times10^{-3}$ cm	x/a	D cm²/sec.	Atmosphere of heating
900	1.0	14.1	6.00	0.235	7.0×10^{-9}	air
	2.0	15.2	5.85	0.260	5.4×10^{-9}	air
	4.0	18.3	5.85	0.313	6.9×10^{-9}	air
800	24.0	18.0	5.85	0.308	7.5×10^{-10}	air
700	4.0	8.2	5.85	0.140	8.9×10^{-11}	air
	5.0	8.9	5.85	0.150	1.2×10^{-10}	hydrogen
	6.0	9.0	5.85	0.154	1.1×10^{-10}	air
	16.0	10.6	5.85	0.180	9.2×10^{-11}	hydrogen
	48.0	14.6	6.20	0.240	8.5×10^{-11}	argon
	140.0	15.0	5.85	0.260	6.0×10^{-11}	argon
600	16.0	6.0	5.85	0.102	5.9×10^{-12}	hydrogen
460	64.0	3.8	5.85	0.065	8.0×10^{-14}	hydrogen

* This work was done on a research program sponsored by the AEC Contract AT-30-1-GEN-367.

[1] J. Groh and G. von Hevesy, Ann. d. Physik **63**, 85 (1920).
[2] G. von Hevesy and A. Obrutcheva, Nature **115**, 674 (1925).
[3] G. von Hevesy and W. Seith, Zeits. f. Physik **56**, 790 (1929).
[4] von Hevesy, Seith and Keil, Zeits. f. Physik **79**, 197 (1932).
[5] W. Seith, Zeits. f. Elektrochemie **39**, 538 (1933).
[6] Steigman, Shockley and Nix, Phys. Rev. **56**, 13 (1939).
[7] W. A. Johnson, Trans. A.I.M.M.E. **143**, 107 (1941).
[8] G. C. Kuczynski, "Self-Diffusion in Sintering of Metallic Particles," J. Inst. Metals, (February, 1949).
[9] G. C. Kuczynski, Phys. Rev. **75**, 1309 (1949).

gradually fills up forming a neck bounded by two cylindrical concave surfaces ABC and $A'B'C$ in Fig. 1.

As in the case of sintering of the spherical metallic particles,[8] self-diffusion is responsible for the formation of this neck. If we assume the vacancy theory of diffusion, then the volume directly underneath the concave surfaces of the neck should have an excess concentration ΔC of holes over the equilibrium concentration C_0. This difference in concentration between the volume underneath the concave surface of the neck and the other parts of the system creates the diffusion gradient necessary for the flow of atoms into the neck area and thus promotes gradual filling up of the latter. The equation of flow can then be written as follows:

$$A\,|(\partial\Delta C/\partial r)_{r=\rho}|\,D'=dV/dt, \qquad (2)$$

where A is the surface through which the vacancies enter the system, V the filled volume between the wire and the plate indicated by shaded area in Fig. 1 $(\partial\Delta C/\partial r)_{r=\rho}$ the gradient of vacancy concentration on the surface of the radius of curvature ρ (such as surface which contour is indicated in Fig. 1 by ABC and $A'B'C$) and D the coefficient of vacancy diffusion. The relationship between the D' and D the coefficient of atomic diffusion is $D'C_0=D$ where C_0 as before designates the equilibrium concentration of vacancies.

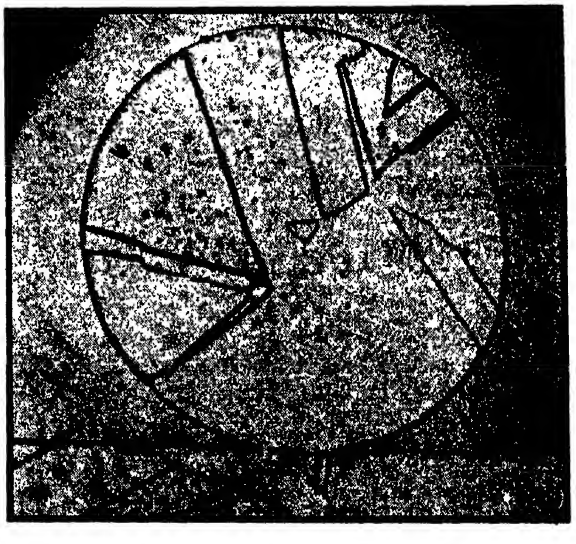
We shall first evaluate the excess hole concentration ΔC. The free energy increase of the system, due to the dissolution of n vacancies raising the concentration from C_0 to $C_0+\Delta C$ is $nkT\ln(C_0+\Delta C/C_0)$ or if ΔC is small $nkT(\Delta C/C_0)$. At the same time, there is the energy decrease due to the decrease of the pressure γ/ρ over the neck area. This energy decrement is approximately $n\delta^3(\gamma/\rho)$ if the diameter of the vacancy is equal to the atomic diameter δ. At equilibrium, these two energies have to be equal which results in the following expression for the increment of concentration under the neck area

$$\Delta C=(\delta^3\gamma/kT\rho)C_0. \qquad (3)$$

This excess concentration of holes causes the diffusion

(a)

(b)



(c)

FIG. 2. (a) Cross section of silver wire sintered to silver cylinder at 800°C for 24 hours, magnified 200 times; (b) same sintered at 900°C for 1 hour, magnified 500 times; (c) same sintered at 700°C for 140 hours, magnified 500 times.

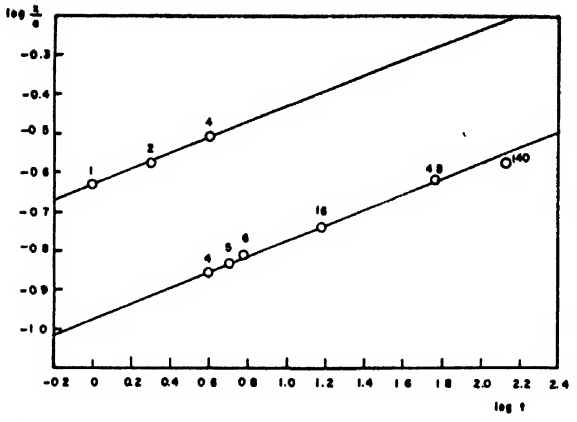


FIG. 3. Log(x/a) *versus* log$t$ plots for silver wires bonded to silver cylinders. The inverse slope of these lines is approximately 5, proving that x^5 is proportional to the time of heating.

of atoms from the inner parts of the system to the neck. When the current is established $\partial C/\partial t=0$ and the concentration will be given by the solution of the Laplace equation

$$\nabla^2 C=0$$

this solution is a function of the radius r drawn from the center of curvature 0 (Fig. 1) of the neck area and the angle θ of this radius with a direction perpendicular to the axis of the wire and parallel to the plate and also passing through the center of curvature. Obviously ρ does not change along the length of the wire. Such a solution can be written in the form

$$C=\sum_{n/0}^{\infty}\frac{C_n\cos n\theta}{r^n},$$

or

$$\Delta C=\sum_{n/1}^{\infty}\frac{C_n\cos n\theta}{r^n}, \tag{4}$$

where

$$\Delta C=C-C_0.$$

Expression (4) should be approximately equal to (3) on the surface of the neck, where $r=\rho$, which implies that $|C_n|$ for $n>1$ are small and corresponding terms can be neglected in our approximation. We may accept, therefore, that

$$\Delta C\cong\delta^3\gamma C_0/kTr$$

and

$$(\partial\Delta C/\partial r)_{r=\rho}=-(\delta^3\gamma C_0/kT\rho^2). \tag{5}$$

Simple geometrical considerations based upon Fig. 1 lead to the following equations which are valid provided $x/a<\cdot 3$

$$\rho=(x^2/4a),\quad A=(\pi x^2L/2a),\quad V=(x^3L/2a),$$

where L is the length of the sintered wire. Inserting these values into (5) and (2) and noting the $D'C_0=D$, we obtain

$$(80\pi/3)(\delta^3\gamma a^2/kTx^4)D=dx/dt,$$

and after integration

$$x^5/a^2=(80\pi/3)(\delta^3\gamma/kT)Dt. \tag{6}$$

Using this equation, we may express the coefficient of diffusion D as a function of the readily measurable variables x, a, T and t.

$$D=3/80\pi(kT/\delta^3\gamma t)x^5/a^2. \tag{7}$$

It has to be noted that in the above derivation, it was tacitly assumed that the contour of the neck ABC in Fig. 1 is semicircular. If it were not so, then there would be a difference of concentration of vacancies proportional to the difference of the inverse radius of the smallest and largest radius of curvature and the diffusion current would tend to equalize the curvature. It is, therefore, reasonable to assume the contour to be circular.

THE EXPERIMENTAL METHOD

As can be seen from the brief introduction, the determination of the coefficient of self-diffusion reduces itself to the sintering of a wire to a large block of the same metal and subsequent measurement of the width $2x$ of the interface formed. In order to demonstrate that in this case the mechanism of sintering is diffusion, the experimental test of Eq. (6) is necessary.

The experiments were performed on very pure silver metal. The sintering of the wire to a flat base was not successful because it was difficult to obtain good contact between the wire and the base. This problem was solved

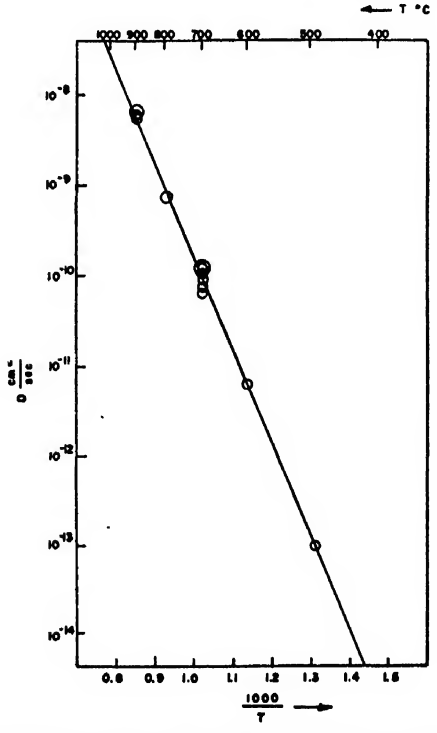


FIG. 4. Variation of volume self-diffusion coefficient of silver with temperature as determined from sintering experiments.

by winding the 0.005 in. diameter silver wire on a cylindrical piece of the same material of about ¼ in. diameter. This diameter was found large enough to allow to neglect the curvature of the cylinder and so consider it as approximately flat. Good care was taken to tighten the wire well on the cylinder without, however, introducing much plastic deformation in the interface. Specimens prepared in this manner were heated in hydrogen, air or argon at 900, 800, 700, 600 and 460°C for various periods of time, as indicated in Table I. After heating, the specimens were nickel plated in order to avoid loosening or deformation of the interfaces during the subsequent metallographic operations. Plated cylinders with wires sintered to them were then mounted in Bakelite and cut, in order to obtain a diametral cross section of the wire in contact with the base.

The results of the measurements of the diameter of the neck, sometimes show large scattering of values, due to more or less perfect contact along the wire. In order to assure maximum accuracy, a great number of readings were taken (more than twenty per temperature-time run) and the most probable diameter was determined from the frequency of occurrence curves.

The plane of the cut was not necessarily perpendicular to the axis of the wire and consequently the width of the interface $2x$ between the wire and the cylinder observed on the cross section was not the true one. This, however, was easily corrected as the apparent distortion of the interface and the known diameter of the wire are in the same proportion as the true interface width and the true wire diameter, known from micrometric measurements. The true width of the interface then was evaluated by multiplying its measured value by the ratio of the true diameter of the wire to its value as measured with the micrometer. Some examples of the polished sections prepared in the manner outlined above are shown in Fig. 2.

RESULTS

Table I lists the half width values of the interface x, along with the radius of wire a, the x/a ratio and the coefficient of self-diffusion D calculated from Eq. (4). The last column contains details of the atmosphere used for each particular heat treatment. It is convenient to convert (6) into logarithmic relationship after dividing both sides by a^5

$$5\log\frac{x}{a}=\log t+\log\left(\frac{80\pi}{3}\frac{\delta^3\gamma D}{kTa^3}\right);$$

the last term is constant at a given temperature. By plotting $\log x/a$ *versus* $\log t$ one can obtain the exponent, which in the case of diffusion flow should be 5, by simple determination of the slope of the obtained straight line.

The $\log x/a/\log t$ plots at 900°C and 700°C are given in Fig. 3. The points fall on the straight lines whose inverse slopes are 5.0 and 5.1 respectively, indicating that the relationship $x^5\sim t$ holds, as should be expected when the diffusion flow is responsible for sintering. The fact that the points corresponding to the heat treatments in three different atmospheres fall on the same straight line, seems to indicate that in the case of silver bonding of wires or particles together is not affected by the surrounding atmosphere, as long as it is neutral. At 700°C, air can be considered neutral to silver.

Figure 4 represents the semilogarithmic plot of $\ln D$ *versus* $1000/T$. The experimental points fall very well on the straight line. The coefficient of self-diffusion of silver as determined from this graph is $D=0.9e^{-45,700/RT}$* and agrees very well with that obtained by W. A. Johnson[7] who used the radioactive isotope method for its determination. Johnson obtained $D=0.89e^{-45,900/RT}$. The value obtained from the experiments of sintering of silver spheres[8] was $D=0.60e^{-45,000/RT}$. As we can see all three values are in close agreement indicating that the wire method may well be used in the determination of the coefficients of self-diffusion of metals. Its greatest advantage is simplicity and great saving of time. In cases like aluminum where the radioactive isotopes are not available the wire method is the only one which permits determination of the self-diffusion coefficient.

ACKNOWLEDGMENTS

The author wishes to express his gratitude to W. E. Kingston, Manager of Sylvania Metallurgical Laboratories, for his support of this project. He also wishes to thank H. Lipson of Sylvania Electric Products, Inc. for his aid in processing the specimens and Mrs. L. Roth for her skillful metallographic work.

* In the calculation of D the value of surface energy for liquid silver, 1000 ergs/cm² was adopted, although it was expected that the surface energy for solid metal should be somewhat higher. However, the error introduced by this choice should not change the order of magnitude of the coefficient of diffusion.

A Graphical Analysis of the Interference Patterns of an Elevated Ultra-High Frequency Antenna under Conditions of Atmospheric Stratification*

F. R. Abbott and C. J. Fisher
U. S. Navy Electronics Laboratory, San Diego, California
(Received September 6, 1949)

Analytical computation of the detailed lobe structure of an elevated meter or centimeter band antenna for a standard atmosphere is quite tedious. In case there are discontinuities in the gradient of refractive index the analysis becomes hopelessly involved.

This paper presents a straightforward method of graphical analysis for the optical region, equally applicable to a standard or stratified atmosphere.

INTRODUCTION

A PICTORIAL presentation of the lobe structure of the low angle radiation pattern from an elevated microwave or v.h.f. antenna is usually constructed with a distorted scale. Horizontal distances may be compressed to a millionth while vertical dimensions are reduced to 1/10,000 or less. Thus elevation angles are greatly exaggerated.

It is advantageous to use such a distorted representation in the construction of ray diagrams. It is very significant that the path difference between intersecting rays under such a transformation is easily measureable and is directly reducible to the path difference of the corresponding rays at full scale.

Thus, if one is concerned with a centimeter or so path length difference between two rays intersecting at a distance of twenty miles, it becomes possible to construct the rays on a distorted scale drawing and measure the path difference on the drawing. The measured path difference multiplied by a constant factor gives the difference for the actual rays and the relative phase is thus determined.

Geometrical path differences on a drawing may be measured to a high degree of accuracy using an ordinary engineer's map measure, the rolling wheel of which has been graduated. It will be shown that the additional term included in the optical path difference may be determined by planimeter measurements of the area enclosed between the rays.

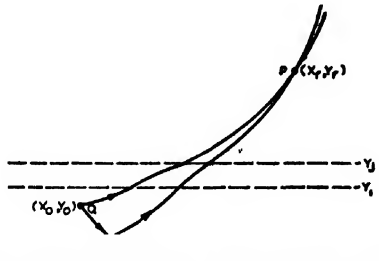


FIG. 1. Two optical paths for radiation proceeding from a source, Q, and traveling through an inversion layer between Y_i and Y_j to a point of intersection, P.

* M. E. Hall of the U. S. Navy Electronics Laboratory Antenna Range staff assisted in the preparation of the material for publication.

THEORY

Consider radiation proceeding from a source at Q to an arbitrary point, P, by two alternate optical paths. The second ray may have been occasioned by a reflection as indicated in Fig. 1 or by other means.

Formally, the optical path length of either ray is given by

$$S=\int_{X_0Y_0}^{X_rY_r} n\,dL, \tag{1}$$

in which n is the refractive index and dL is the element of geometrical path length which, in turn, is given by

$$dL=[1+(dY/dX)^2]^{\frac{1}{2}}dX \cong[1+\tfrac{1}{2}(dY/dX)^2-\tfrac{1}{8}(dY/dX)^4+\cdots]dX. \tag{2}$$

The third and all ensuing terms in the expression for dL are very small for rays at low vertical angles and will henceforth be disregarded. The refractive index of the lower atmosphere, designated by n, is assumed to be dependent only on the elevation. It is further assumed that the vertical refractive index curve, $n=f(Y)$, may be replaced by a series of sections over which dn/dY is constant. n is expressible as

$$n=1+q(Y), \tag{3}$$

in which $q(Y)$ is a small fraction, generally 0.0004 or less for the case considered.[1] The radius of curvature of rays in a medium of slowly changing refractive index is given by[2]

$$1/R=(d/dY)\log n=(1/n)(dn/dY). \tag{4}$$

For values of n very close to unity

$$1/R\cong(dn/dY). \tag{5}$$

The two rays in Fig. 1 thus proceed through three zones. In the lower zone the refractive index is increasing with increased altitude and since a ray bends into the denser medium, the curvature is upward. From elevations Y_i to Y_j the refractive index is decreasing and the

[1] I. Katz and J. M. Austin, "Qualitative survey of meteorological factors affecting microwave propagation," M.I.T. Radiation Lab. Report No. 488 (June 1, 1944).

[2] P. K. L. Drude, *The Theory of Optics* (Longmans, Green and Company, New York, 1902), p. 306.

curvature is downward. Above Y_j the refractive index increases and the curvature is again upward.

Designate the quantities associated with the two rays by the subscripts a and b, respectively. The path lengths from Eqs. (1)–(3) become

$$S_a \approx \int_{X_0}^{X_r} [1+q(Y)_a][1+\tfrac{1}{2}(dY/dX)_a^2]dX, \quad (6)$$

and

$$S_b \approx \int_{X_0}^{X_r} [1+q(Y)_b][1+\tfrac{1}{2}(dY/dX)_b^2]dX. \quad (7)$$

By neglecting terms of the order $\frac{1}{2}q(Y)(dY/dX)^2$ and higher, the optical path difference becomes

$$S_b-S_a \approx \int_{X_0}^{X_r} [q(Y)_b-q(Y)_a + \tfrac{1}{2}(dY/dX)_b^2-\tfrac{1}{2}(dY/dX)_a^2]dX. \quad (8)$$

Designating the vertical gradient of refractive index by $n'=dn/dy$, we have

$$q(Y)_b-q(Y)_a=n_b-n_a=\int_{Y_a}^{Y_b} n'dY,$$

and

$$S_b-S_a=\tfrac{1}{2}\int_{X_0}^{X_r} [(dY/dX)_b^2-(dY/dX)_a^2]dX + \int_{X_0}^{X_r}\int_{Y_a}^{Y_b} n'dYdX. \quad (9)$$

The first integral in Eq. (9) is simply the geometrical path length difference for the two rays and is expressible as L_b-L_a.

Since the gradient of refractive index, n', is constant except for discontinuities at $Y=Y_i$ and $Y=Y_j$ the double integral is equivalent to the sum of the three area integrals, each multiplied by the value of n' for the area concerned. This is expressible by $\sum_i n_i'A_i$ in which A_i is the area characterized by a refractive index gradient n_i' and bounded by the paths a and b and the elevation levels Y_i and/or Y_j.

Thus, the optical path difference becomes

$$S_b-S_a=L_b-L_a+\sum n_i'A_i. \quad (10)$$

The construction of a ray diagram involves the reduction in dimensions expressed by

$$dy=CdY, \qquad dx=DdX, \quad (11)$$

in which C and D assume values of the order of 1/10,000

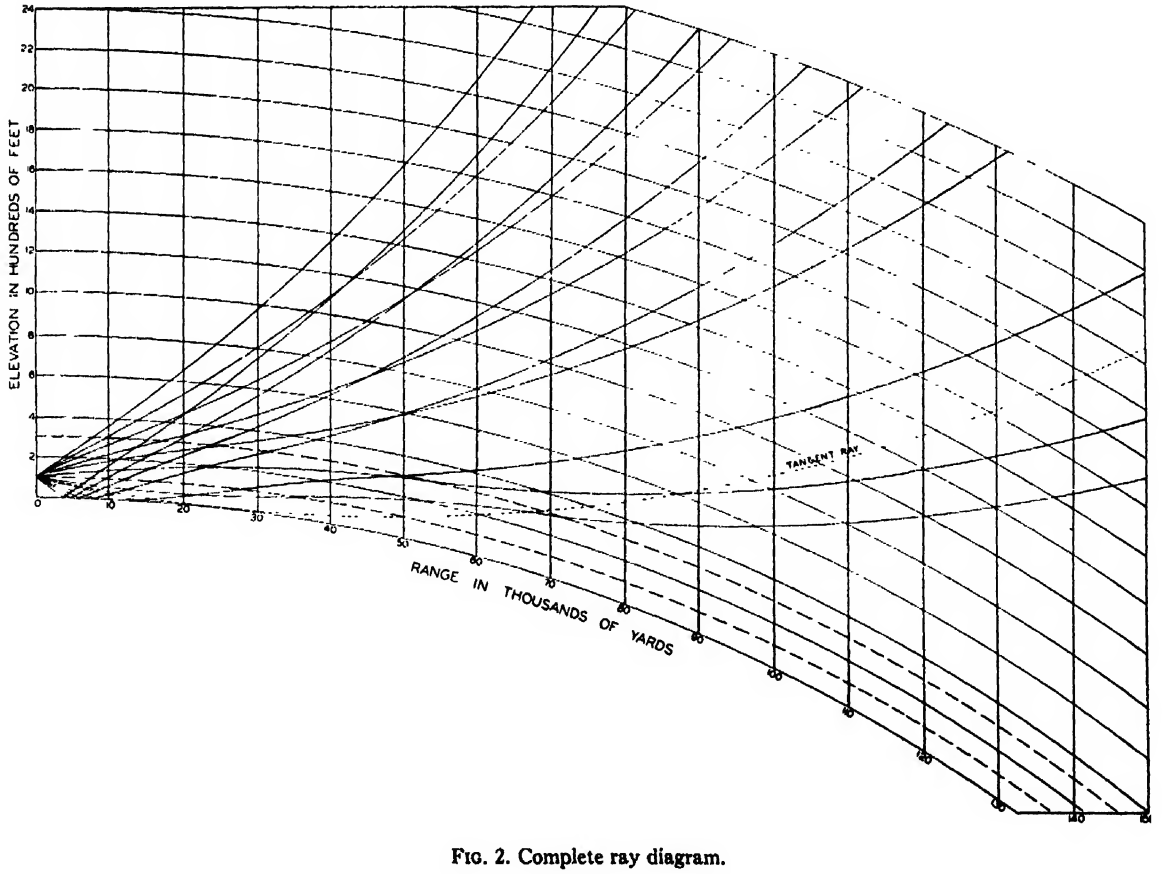


FIG. 2. Complete ray diagram.

and 1/1,000,000, respectively. In the scale coordinates Eq. (8) becomes

$$s_b - s_a = \frac{D}{2C^2}\int_{X_0}^{X_r}[(dy/dx)_b^2 - (dy/dx)_a^2]dx + \frac{1}{CD}\int_{X_0}^{X_r}\int_{Y_a}^{Y_b} n'dydx, \quad (12)$$

which like (9) is expressible as

$$s_b - s_a = D/C^2(l_b - l_a) + 1/CD\sum_i n_i'\alpha_i, \quad (13)$$

where α and l are the reduced scale area and length symbols.

The radius of curvature of the rays on the reduced scale drawing is proportional to the radius of curvature of the actual rays. The proportionality factor is determined as follows:

From elementary calculus the radius of curvature of the actual rays is given by

$$R = \frac{[1+(dY/dX)^2]^{\frac{3}{2}}}{d^2Y/dX^2}.$$

For rays near grazing incidence $(dY/dX)^2$ is very small so that

$$1/R \approx d^2Y/dX^2. \quad (14)$$

The reduced scale radius, r, follows from the transformation of Eqs. (11) applied to (14) and is given by

$$1/r \approx C/D^2 d^2Y/dX^2. \quad (15)$$

By Eq. (5), Eq. (15) becomes

$$1/r \approx (C/D^2)n'. \quad (15a)$$

Introducing r_i, the scale radius of curvature of the rays, into Eq. (13) we have for the path difference

$$s_b - s_a = (D/C^2)[(l_b - l_a) + \sum_i \alpha_i/r_i]. \quad (16)$$

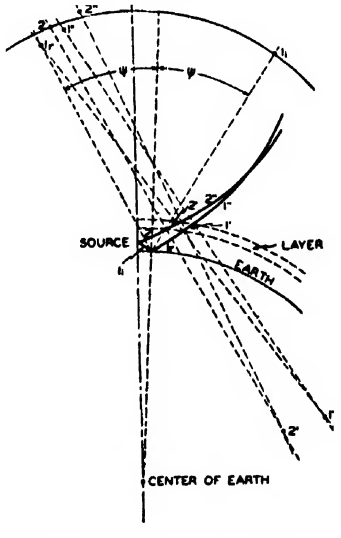


FIG. 3. Illustration of method of construction of a ray diagram.

As is shown in the Appendix the modification of the actual index of refraction curve does not alter the total path length difference of any pair of intersecting rays. It has been shown[3] that to construct a diagram in which the rays in regions of standard refraction are represented by straight lines requires the use of an effective earth's radius equal to 4/3 the actual radius of the earth. This defines the slope of a ray passing from the source through any point.

Since it is imperative to keep the angle of inclination of rays and the scaled earth surface small in order that the neglected terms of Eq. (2) do not introduce appreciable error, a curve of modified index has been selected that will minimize both. This curve is such that the resultant curvature of both the surface of the earth and the rays in regions of standard refraction is equal to half the curvature of the earth's surface relative to a ray in standard atmosphere obtained from the above relation. This gives a radius of curvature for both rays in the region of standard refraction and the surface of the earth equal to 8/3 the actual radius of the earth. Thus, from Eq. (15) the scaled earth radius equals $(D^2/C)R_e$, where R_e is the modified radius of earth curvature and is equal to 8/3 actual earth radius.

In the actual construction of a ray diagram appropriate values of the scale factors, C and D, are chosen and the earth's surface is drawn using the resulting value for the scaled earth radius as given above. The diagram shown in Fig. 2 was drawn on a chart 27×40 in. on which one inch vertical represents 150 ft. and one inch horizontal represents 2 miles. A layer of inversion of refractive index is indicated as being between the dashed lines. The transformation constants, C and D, assume values of 1/1800 and 1/144,000, respectively and result in a scaled earth radius of 58 in.

CONSTRUCTION OF RAY DIAGRAM

Using the location of the source of radiation as center, an arc of scaled earth radius is struck above the line representing the earth's surface. Equally spaced points along this arc provide centers of curvature for equally spaced rays emanating from the source. Consider a direct and reflected ray whose centers, for the sake of explanation of construction, shall be the arbitrary points, 2 and 1_i, respectively, as shown in Fig. 3. When a direct ray, such as ray No. 2, is incident on the lower boundary of an inversion layer the curvature reverses in sign. The new center lies on an extension of the line passing through the center of curvature of the incident ray and the point of incidence. The radius of curvature of the ray while passing through the inversion layer is again defined by Eq. (15) and can be found from the modified index curve. This procedure is repeated when the ray meets the upper boundary of the layer and the new center is found on an extension of the second radius

[3] Schelleng, Burrows, and Ferrell, "Ultra-short-wave propagation," Proc. I.R.E. 21, 427 (1933).

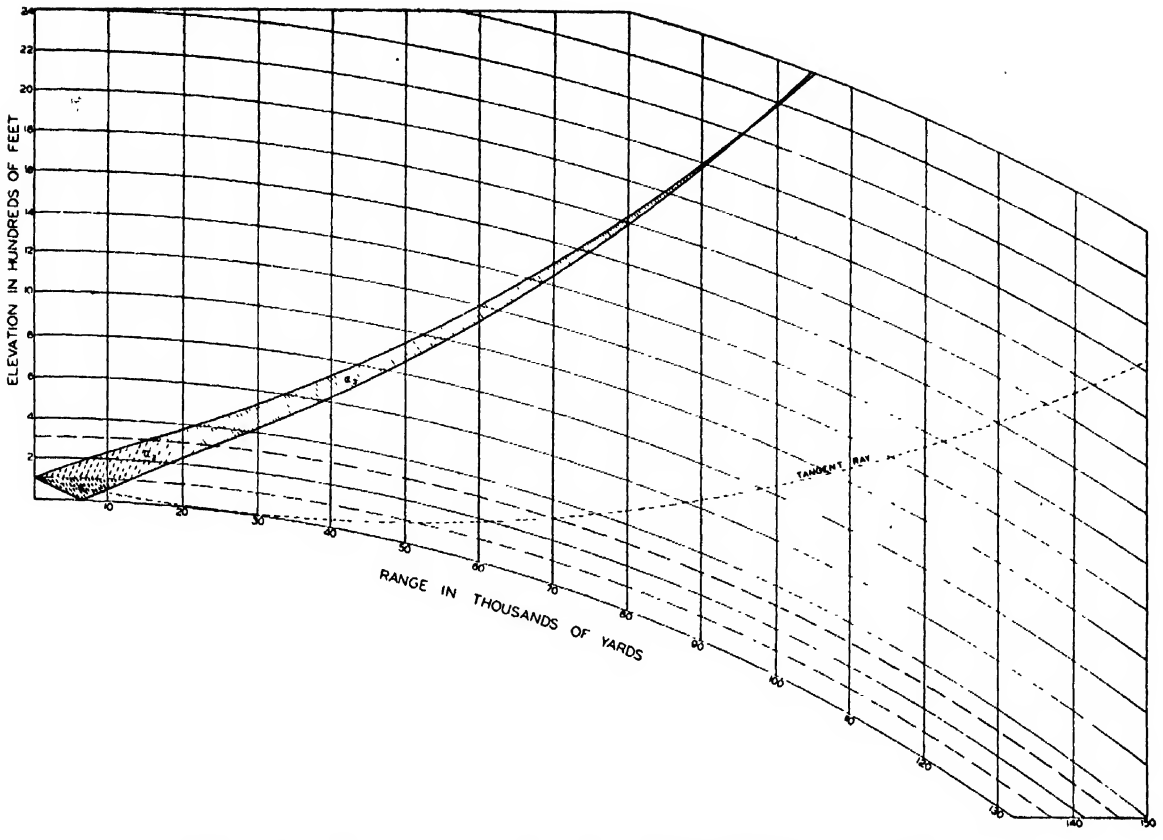


FIG. 4. Areas, α_i, used in evaluating the resultant field, E, of Eq. (17) at any point of intersection.

passing through the point of incidence at the top of the layer. This is shown as line 2′2″ in Fig. 3.

In the case of a ray reflected from the earth's surface such as ray No. 1 in Fig. 3, an extension of the earth's radius is drawn through the point of reflection. Since the angle of reflection is equal to the angle of incidence, the center of curvature of the reflected ray is found on a line making an angle, ψ, with this extension of the earth's radius. The reflected ray from this point on is constructed exactly as the direct ray given above.

The areas, α_i, of Eq. (16) bounded by the two rays and the levels of discontinuity in refractive index gradient shown in Fig. 4 are measured with a planimeter. The geometrical path difference is measured by running a map measure forward over one ray and backward along the alternate ray to the starting point. It should be remembered that terms of α_i/r_i are added algebraically and due regard should be given to change of sign of r_i when the direction of curvature changes.

It is essential to select values of C and D such that the slope of the rays remains small. Otherwise the disregarded terms of Eq. (2) become excessively large.

The intensity of each component of the field normally decreases with distance in accordance with the inverse distance law of radiation fields. The two factors necessary to account for such variation are designated by $1/(d)^{\frac{1}{2}}$ and $1/(\delta)^{\frac{1}{2}}$ where δ is the distance between individual rays of the ray system at a distance, d, from the transmitter.

We then finally obtain the resultant field, E, at any point of intersection of two rays with reference to E_0, the field at some reference distance, d_0, in the form

$$\frac{E}{E_0}=\frac{1}{(d)^{\frac{1}{2}}}\left\{\frac{1}{(\delta_a)^{\frac{1}{2}}}+\frac{\rho}{(\delta_b)^{\frac{1}{2}}}\exp j\beta\left[\phi+\frac{D}{C^2}\left(l_b-l_a+\Sigma_i\frac{\alpha_i}{r_i}\right)\right]\right\}, \quad (17)$$

where $\beta=2\pi/\lambda$, δ_0 is the ray spacing at distance d_0 from the source, usually taken as one meter, and the subscripts, a and b, refer to the direct and reflected rays, respectively. Equation (17) is valid so long as $1/(\delta)^{\frac{1}{2}}$ is finite.

The phase shift, ϕ, and the magnitude, ρ, are introduced to allow for phase rotation and change of intensity of the lower ray on reflection. In the case of horizontally polarized waves propagated over sea water, ϕ goes to $\lambda/2$, and ρ becomes unity.

Figure 5 is a plot of the field pattern of a 3000-Mc source elevated 100 ft. above the surface of the earth. It was secured by evaluating Eq. (17) at all the points of intersection of rays on the corresponding ray diagram.

The foregoing method of analysis presents an unique

and simple procedure for computation of the radiation field in the optical zone, for an elevated very high frequency antenna in any type of stratified atmosphere. Good correlation has been obtained between the measured field in the optical zone under conditions of atmospheric stratification and the calculated field for the same condition.

A complete field analysis below the optical horizon must include diffracted, scattered and partially reflected components of radiation[4] none of which is considered in the above analysis.

A number of predicted low level coverage diagrams similar to Fig. 5 are available in reports WP-11, WP-14, and WP-15 of the U. S. Navy Radio and Sound Laboratory.[5] These include patterns of 200- and 3000-mcs sources for 13 types of non-linear atmospheric refractive index gradients.

APPENDIX. PROOF OF THE INVARIANCE OF OPTICAL PATH DIFFERENCE TO THE CURVATURE IMPOSED ON SPACE BY THE USE OF A MODIFIED INDEX OF REFRACTION

There are a variety of forms that the modified index of refraction curve may take as is shown in Fig. 6. Each one lends more or less convenience to a given application. It would be well to ask what effect the modification of the gradient of the index of refraction has on the analysis presented.

The path difference for the two rays depicted in Fig. 1 is given by an expression of the form

$$S_b - S_a = (L_b - L_a) + \Sigma_i A_i/R_i,$$

where $L_b - L_a$ is the geometric path length difference and A_i is the area bounded by the rays in the i region. R_i is the radius of curvature of the ray in the i region.

For values of n close to unity,

$$1/R \approx dn/dY \approx d^2Y/dX^2,$$

which leads to

$$S_b - S_a = (L_b - L_a) + \Sigma_i n' A_i. \tag{18}$$

From Eq. (8) we see that the geometric path length difference is expressible as

$$L_b - L_a = \tfrac{1}{2}\int [(dY/dX)_b^2 - (dY/dX)_a^2]dX$$

$$= \tfrac{1}{2}\int [(Y_b')^2 - (Y_a')^2]dX. \tag{19}$$

Let z_a be the slope of the reflected ray at the point of reflection and z_b be the slope of the direct ray at the point of reflection. Then

$$Y_a' = z_a - \int_0^X n_a' dX, \tag{20}$$

$$Y_b' = z_b - \int_0^X n_b' dX. \tag{20a}$$

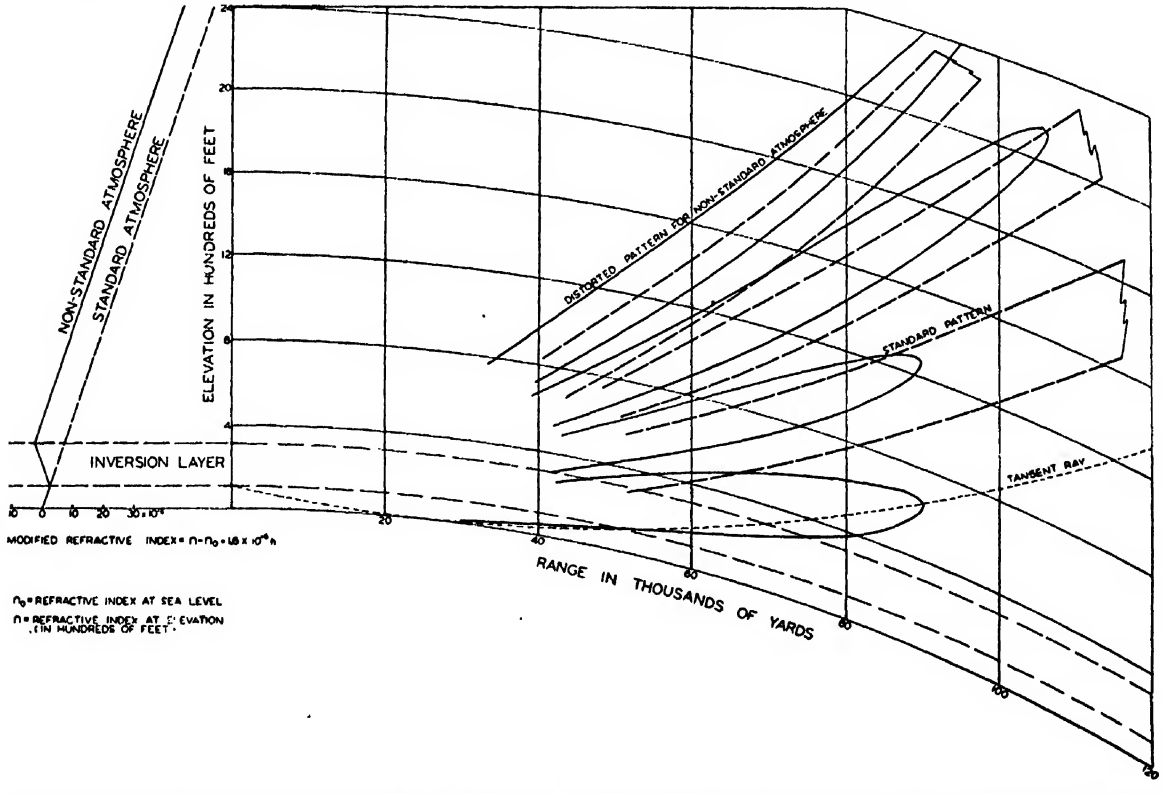


FIG. 5. Plot of a field pattern showing low level coverage for a modified refractive index equal to $n - n_0 + 1.8 \times 10^{-6}h$, where n_0 is the refractive index at sea level and n is the refractive index at elevation h. (h is measured in hundreds of feet.)

[4] J. B. Smith and L. G. Trolese, "Propagation of radio waves in the lower troposphere," Proc. I.R.E. 35, No. 11, 1198–1202 (November, 1947).

[5] U. S. N. R. and S. Lab. Report WP-11, "Procedure and charts for estimating low level coverage of shipboard 200 mcs radars under conditions of pronounced refraction" (April 10, 1944). U. S. N. R. and S. Lab. Report WP-14, "Predicted low level coverage of S-band radars as affected by weather" (November 1, 1944). U. S. N. R. and S. Lab. Report WP-15, "Predicted low level coverage of 200 mcs radars as affected by weather" (November 4, 1944).

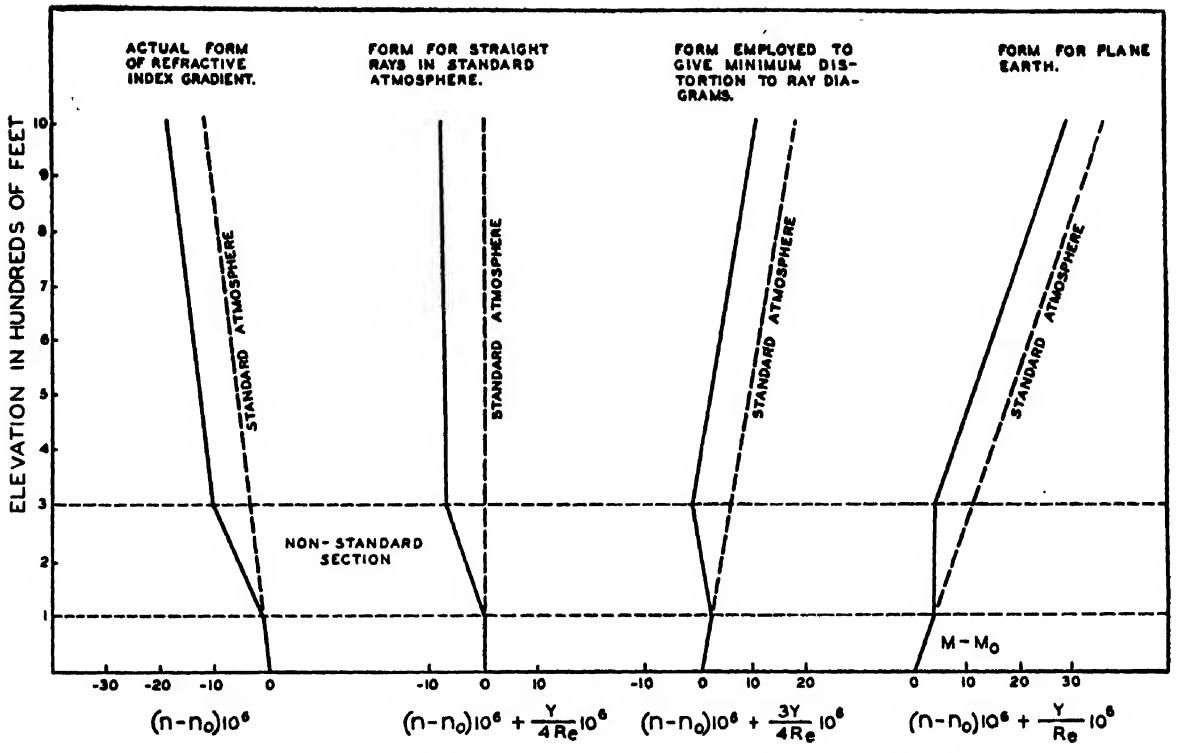


FIG. 6. Four useful types of representation of atmospheric refractive index gradient. All represent the same condition.

Let n' be modified by an amount, k, and designate the new slopes by U_a' and U_b'. It is seen from Fig. 6 that the factor, k, assumes values of $(Y/R_e)10^6$, $(Y/4R_e)10^6$ etc., where Y is the elevation in hundreds of feet and R_e is the radius of the earth.

$$U_a' = z_a - \int_0^X (n_a' + k)dX, \tag{21}$$

$$U_b' = z_b - \int^X (n_b' + k)dX. \tag{21a}$$

From (21) and (21a)

$$U_b' - U_a' = z_b - z_a - \int^X (n_b' - n_a')dX = Y_b' - Y_a', \tag{22}$$

$$U_b' + U_a' = z_b + z_a - \int_0^X (n_b' + n_a')dX - 2k\int_0^X dX = Y_b' + Y_a' - 2kX. \tag{23}$$

Hence,

$$(U_b')^2 - (U_a')^2 = (U_b' - U_a')(U_b' + U_a') = (Y_b')^2 - (Y_a')^2 - (Y_b' - Y_a')2kX. \tag{24}$$

Let Δ_g be the increment in geometric path length difference that results from modifying n' by the factor k. From Eq. (19)

$$\Delta_g = \tfrac{1}{2}\left\{\int[(U_b')^2 - (U_a')^2]dX - \int[(Y_b')^2 - (Y_a')^2]dX\right\}. \tag{25}$$

By substituting (24) in (25) we see that

$$\Delta_g = -k\int_Q^P (Y_b' - Y_a')XdX. \tag{26}$$

Integrating Eq. (26) by parts and letting

$$u = X; \quad du = dX,$$

and

$$dv = (Y_b' - Y_a')dX; \quad v = Y_b - Y_a,$$

we see that

$$\Delta_g = -k\left[X(Y_b - Y_a)\Big|_Q^P - \int_Q^P (Y_b - Y_a)dX\right]. \tag{27}$$

In the first term, $Y_b - Y_a$ is equal to zero at the points P and Q since the rays cross and the difference in height is zero. The integral term represents the area, A, bounded by the two rays. Therefore, $\Delta_g = kA$.

From Eq. (18), the purely optical contribution to path difference is given by the term $\Sigma_i n_i' A_i$. By modifying the refractive index gradient by a factor, k, this term takes the form $\Sigma_i (n_i' + k)A_i$.

Let Δ_0 be the variation in the purely optical component of the path length difference of two rays passing through media having refractive indices n and $(n+k)$, respectively. Then

$$\Delta_0 = \Sigma_i n_i' A_i - \Sigma_i (n_i' + k)A_i = -\Sigma_i kA_i.$$

Since $\Sigma_i A_i$ represents the total area, $\Delta_0 = -kA$.

Thus, the total change in path length difference observed when n is modified by a factor, k, is

$$\Delta_t = \Delta_g + \Delta_0 = kA - kA = 0.$$

It is important to note that the relative total path length difference is independent of the modification of the gradient of the index of refraction.

It is elementary to prove that the elevation and distance to the point of intersection of two rays is also independent of the choice of modified index.

The Determination of the Impulsive Response of Variable Networks

LOTFI A. ZADEH
Columbia University, New York, New York
(Received November 21, 1949)

Starting with the differential equation relating the output and the input of a linear varying-parameter network, it is shown that the impulsive response of the system is related to a Green's function associated with the system through a linear operator which is the adjoint of the right-hand operator in the given differential equation. A perturbation procedure for the determination of the impulsive response of a slowly varying network is outlined. Use of the method is illustrated by a simple example involving a band width-modulated RC half-section.

GENERAL THEORY

THE term *variable network* is commonly used to describe a network in which one or more element-values are dependent upon a combination of the three variables—time, input, and output. The variability of transmission characteristics of variable networks makes them ideally suited for many applications in the fields of communication and control. In practice, however, the use of variable networks has been, and still is, very limited mainly because of the lack of adequate means for their analysis, synthesis and mechanization.

The most basic problem in the theory of variable networks is that of analysis. In general, the analysis of a variable network in which the element-values are functionals of the input, output, and time, presents mathematical problems of a prohibitive degree of complexity—problems which for all practical purposes are unsolvable except through use of high speed automatic computers. There are, however, certain classes of variable networks which are amenable to purely analytical treatment. The most important of these is the class of so-called linear varying-parameter networks, that is, linear networks in which the element-values are functions of time only, and are independent of both the input and the output. Familiar examples of such networks are the amplitude and frequency modulators and—as a limiting case—the ordinary fixed linear networks. To simplify the terminology, the term *variable network* as used in this paper will refer to a linear varying-parameter network.

The behavior of a variable network may be described in several different ways—some of which are implicit and some explicit. Usually, the behavior is described *implicitly* by means of a differential equation relating the output of the network, $e_2(t)$, to the input $e_1(t)$. This equation is, in general, of the form

$$[a_n(t)p^n+\cdots+a_1(t)p+a_0(t)]e_2(t) = [b_m(t)p^m+\cdots+b_1(t)p+b_0(t)]e_1(t), \quad (1)$$

where $p=d/dt$ and the a's and b's are known functions of time. Equation (1), which is sometimes referred to as the *fundamental* equation of the system, may be written more compactly as

$$L(p;t)e_2(t)=K(p;t)e_1(t), \quad (2)$$

where $L(p;t)$ and $K(p;t)$ stand, respectively, for the left-hand and right-hand operators in Eq. (1). The input $e_1(t)$ is usually assumed to be zero for $t<0$, and the network itself is assumed to be unexcited, though not necessarily fixed, prior to $t=0$. Thus, the problem of the determination of the response to a specified input $e_1(t)$ reduces to the solution of the differential equation represented by Eq. (1), subject to homogeneous initial conditions at $t=0-$.

A commonly used form of *explicit* description of the behavior of a variable network is based on the notion of the *impulsive response* of the network. The impulsive response $W(t,\xi)$ is defined as the response of the system at $t=t$ to a unit impulse applied at $t=\xi$. In other words, $W(t,\xi)$ is the solution of the equation

$$L(p;t)W(t,\xi)=K(p;t)\delta(t-\xi), \quad (3)$$

where $\delta(t-\xi)$ represents a unit impulse at $t=\xi$. The salient property of $W(t,\xi)$ is that the response to any given input $e_1(t)$, may be expressed in terms of $W(t,\xi)$ and $e_1(t)$ through the superposition integral

$$e_2(t)=\int_{-\infty}^{\infty} W(t,\xi)e_1(\xi)d\xi. \quad (4)$$

Thus, when the behavior of a network is described by means of its impulsive response, the determination of the response of the network to any given input reduces to performing the relatively simple operation expressed by Eq. (4).

A more convenient form of explicit description of the behavior of a variable network is based on the notion of the *system function*[1] of the network. Briefly, the system function of a variable network is defined as a function $H(j\omega;t)$ such that $H(j\omega;t)e^{j\omega t}$ represents the steady-state response of the network to $e_1(t)=e^{j\omega t}$. In other words, the system function $H(j\omega;t)$ is the complex envelope of the response of the network to $e^{j\omega t}$. It is shown in [1] that $H(j\omega;t)$ is the steady-state solution of the differential equation

$$L(p+j\omega;t)H(j\omega;t)=K(j\omega;t), \quad (5)$$

and that $H(j\omega;t)$ and $W(t,\xi)$ are related to each other

[1] L. A. Zadeh, "Frequency analysis of variable networks," Proc. I.R.E. 38, 291–299 (1950).

by the relations:

$$H(j\omega; t)=\int_{-\infty}^{\infty} W(t, \xi)e^{-j\omega(t-\xi)}d\xi \tag{6}$$

and

$$W(t, \xi)=\frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega; t)e^{j\omega(t-\xi)}d\omega. \tag{7}$$

The salient characteristic of $H(j\omega; t)$ is that the response to any given input $e_1(t)$ may be expressed in terms of $H(j\omega; t)$ and $e_1(t)$ through the relation

$$e_2(t)=\mathfrak{L}^{-1}\{H(s; t)E_1(s)\}, \tag{8}$$

where $E_1(s)$ is the Laplace transform of $e_1(t)$ and $\mathfrak{L}^{-1}$ represents the inverse Laplace transformation. In evaluating the inverse Laplace transform of $H(s; t)E_1(s)$, t should be treated as if it were a constant parameter. Thus, when the behavior of a variable network is described by means of its system function $H(j\omega; t)$, the determination of the response of the network to any given input requires only the use of a standard table of Laplace transforms. This constitutes one of the advantages of describing the behavior of a variable network by its system function $H(j\omega; t)$ rather than by its impulsive response $W(t, \xi)$.

The central problem in the analysis of a variable network is that of obtaining an explicit description of the behavior of the network starting from an implicit description of the behavior. In other words, the problem is to determine either the impulsive response $W(t, \xi)$ or the system function $H(j\omega; t)$ from the knowledge of the fundamental equation of the system. In general, this problem cannot be solved exactly, so that in most cases it is necessary to use approximate methods for the determination of $W(t, \xi)$ or $H(j\omega; t)$. The problem of the determination of $H(j\omega; t)$ is treated in [1]. The problem to be considered in the following is that of the determination of $W(t, \xi)$ in the case of slowly varying networks. In such networks the coefficients $a_\nu(t)$ and $b_\mu(t)$ in Eq. (1) do not vary appreciably over the effective duration[2] of the impulsive response of the network. Most of the variable networks encountered in practice belong to this category.

DETERMINATION OF $W(t, \xi)$

From the definition of the impulsive response of a variable network it follows that $W(t, \xi)$ is the solution of Eq. (3) subject to homogeneous initial conditions at $t=\xi-$. On the other hand, the commonly used Green's function $G(t, \xi)$, which is associated with Eq. (2), is in effect the impulsive response of a system whose fundamental equation is

$$L(p; t)e_2(t)=e_1(t). \tag{9}$$

Stated more precisely, $G(t, \xi)$ is the solution of the differential equation

$$L(p; t)G(t, \xi)=\delta(t-\xi), \tag{10}$$

subject to homogeneous initial conditions at $t=\xi-$, or equivalently, to the following conditions:

$$G^{(\lambda)}(t, \xi)_{t=\xi+}=0, \quad \lambda=0, 1, 2, \cdots, n-2,$$
$$G^{(n-1)}(t, \xi)_{t=\xi+}=\frac{1}{a_n(\xi)}, \tag{11}$$

at $t=\xi+$. In view of Eqs. (3) and (10) it is evident that $W(t, \xi)$ and $G(t, \xi)$ are closely related to each other. To obtain the form of this relation it is sufficient to apply the superposition integral to Eq. (9), treating $K(p; t)\delta(t-\xi)$ as the input to the system described by Eq. (9). Thus, we find

$$W(t, \xi)=\int_{-\infty}^{\infty} G(t, u)[K(p; u)\delta(u-\xi)]du, \tag{12}$$

where u is the variable of integration and $p=d/du$. Making use of the identity

$$\int_{-\infty}^{\infty} f(u)p^k\delta(u-\xi)du\equiv(-1)^k f^{(k)}(\xi), \tag{13}$$

Eq. (12) reduces to the desired relation, namely,

$$W(t, \xi)=b_0(\xi)G(t, \xi)-\frac{\partial}{\partial\xi}[b_1(\xi)G(t, \xi)]$$
$$+\cdots+(-1)^m\frac{\partial^m}{\partial\xi^m}[b_m(\xi)G(t, \xi)]. \tag{14}$$

This relation may be expressed more compactly in the form

$$W(t, \xi)=K^*(p; \xi)G(t, \xi) \tag{15}$$

where p should be interpreted as $d/d\xi$, and the operator $K^*(p; \xi)$ is the adjoint of $K(p; \xi)$.

Equation (15) shows that the problem of the determination of $W(t, \xi)$ may be reduced to the somewhat simpler problem of finding a Green's function satisfying the initial conditions expressed by Eq. (11). This, evidently, is equivalent to finding the impulsive response of the system described by Eq. (9).

In finding $G(t, \xi)$ we shall make use of the assumption that the operator $L(p; t)$ does not vary appreciably over the effective duration of $G(t, \xi)$. In general, Eq. (10) may be rewritten as

$$L(p; \xi)G(t, \xi)=\delta(t-\xi)+[L(p; \xi)-L(p; t)]G(t, \xi), \tag{16}$$

where the coefficients in $L(p; \xi)$ are the constants $a_\nu(\xi)$ and $b_\mu(\xi)$, which are the values of $a_\nu(t)$ and $b_\mu(t)$ at the instant of application of the unit impulse. Now since the network has been assumed to be a slowly varying one, $a_\nu(t)$ and $b_\mu(t)$ will not differ appreciably from $a_\nu(\xi)$ and

[2] The effective duration of a signal is, roughly speaking, the length of the interval outside of which the amplitude of the signal is small by comparison with its maximum value.

$b_\mu(\xi)$ over the effective duration of $G(t, \xi)$. Hence, the second term in the right-hand member of Eq. (16) may be regarded as a perturbation and accordingly a perturbation procedure[3] may be used for the determination of $G(t, \xi)$.

As the first step in the procedure, $G(t, \xi)$ is expressed in the form of a perturbation series

$$G(t, \xi)=G_1(t, \xi)+G_2(t, \xi)+G_3(t, \xi)+\cdots, \quad (17)$$

where $G_1(t, \xi)$ is the first approximation to $G(t, \xi)$, and $G_2(t, \xi)$, $G_3(t, \xi)$, etc., are the successive correction terms. Substituting this series into Eq. (16), it is found that Eq. (16) reduces to a set of iterative relations

$$L(p; \xi)G_1(t, \xi)=\delta(t-\xi), \quad (18)$$

$$L(p; \xi)G_{k+1}(t, \xi) = [L(p; \xi)-L(p; t)]G_k(t, \xi), \quad k=1, 2, \cdots, \quad (19)$$

which formally provide the solution of the problem.

The formal solution represented by Eqs. (18) and (19) may be simplified to a considerable extent, first, by making further use of the assumption that $L(p; t)$ is a slowly varying function of time, and second, by expressing Eqs. (18) and (19) in Laplace transform form. Thus, expanding the right-hand operator in Eq. (19) in a Taylor series around $t=\xi$ and neglecting all but the first, say, N terms in this expansion and also in Eq. (17),[4] the iterative relation expressed by Eq. (19) assumes the following form:

$$-L(p; \xi)G_{k+1}(t, \xi)\cong\left[(t-\xi)\frac{\partial L(p; \xi)}{\partial \xi}+\frac{1}{2!}(t-\xi)^2\frac{\partial^2 L(p; \xi)}{\partial \xi^2}+\cdots+\frac{1}{N!}(t-\xi)^N\frac{\partial^N L(p; \xi)}{\partial \xi^N}\right]G_k(t, \xi). \quad (20)$$

Next, transforming both sides of Eq. (20) with respect to $t'=t-\xi$ and letting $F_k(s; \xi)$ (or simply F_k) represent the Laplace transform of $G_k(t, \xi)$, we obtain:

$$F_1=1/L(s; \xi) \quad (21)$$

and

$$F_{k+1}\cong F_1\left\{\frac{\partial}{\partial s}\left[F_k\frac{\partial L(s; \xi)}{\partial \xi}\right]+\cdots+\frac{(-1)^{N-1}}{N!}\times\frac{\partial^N}{\partial s^N}\left[F_k\frac{\partial^N L(s; \xi)}{\partial \xi^N}\right]\right\}, \quad k=1, 2, \cdots. \quad (22)$$

The above relations express essentially the solution of the problem. It will be noted that from these relations the terms F_1, F_2, F_3, etc. can be obtained in succession by using nothing more complicated than routine differentiation. Once F_1, F_2, F_3, etc. have been determined, the terms $G_1(t, \xi)$, $G_2(t, \xi)$, $\cdots$, $G_N(t, \xi)$, forming the Nth order approximation to $G(t, \xi)$, can be found by using a table of Laplace transforms. And finally, the Nth order approximation to $W(t, \xi)$ is obtained in the form

$$W(t, \xi)\cong W_1(t, \xi)+W_2(t, \xi)+\cdots+W_N(t, \xi), \quad (23)$$

by substituting the Nth order approximation to $G(t, \xi)$ into Eq. (15). These, in brief, are the steps leading to the determination of $W(t, \xi)$.

In most practical cases Eq. (23) develops as a power series in a small parameter. Thus the questions of convergence and the accuracy of the obtained approximation to $W(t, \xi)$ can usually be resolved without significant difficulties. Actually, many if not most of the variable networks encountered in practice are such that the first two or three terms in Eq. (23) provide an adequate approximation to $W(t, \xi)$.

The use of the foregoing method is well illustrated by a simple example for which an exact solution can be obtained by classical methods. Consider a low pass *RC* half-section with a periodically varying band width. The behavior of such a network is described by the first-order differential equation

$$(p+1+\epsilon\cos\omega_0 t)e_2(t)=(1+\epsilon\cos\omega_0 t)e_1(t), \quad (24)$$

where

$$\epsilon=\frac{\text{amplitude of band width variation}}{\text{mean band width}},$$

$$\omega_0=\frac{\text{frequency of band width variation}}{\text{mean band width}},$$

$$t=\text{actual time}\times\text{mean band width}.$$

Using the standard method of solution of first-order equations, the impulsive response of the network is found to be

$$W(t, \xi)=(1+\epsilon\cos\omega_0\xi)\times\exp[-(t-\xi)-(\epsilon/\omega_0)(\sin\omega_0 t-\sin\omega_0\xi)]. \quad (25)$$

Using the approximate method, we find

$$F_1=\frac{1}{L(s; \xi)}=\frac{1}{s+1+\epsilon\cos\omega_0\xi} \quad (26)$$

and

$$F_2=\frac{\epsilon\omega_0\sin\omega_0\xi}{(s+1+\epsilon\cos\omega_0\xi)^3}, \quad (27)$$

from which

$$W(t, \xi)\cong(1+\epsilon\cos\omega_0\xi)[1+(\epsilon\omega_0/2)(t-\xi)^2\sin\omega_0\xi]\times\exp[-(t-\xi)(1+\epsilon\cos\omega_0\xi)]. \quad (28)$$

[3] The perturbation method used in the sequel is based on Schelkunoff's wave perturbation method, "Approximate solution of linear differential equations," Bell Sys. Tech. J. 27, 350–364 (1948).

[4] In some cases the number of terms conserved in the two expansions need not be the same.

It can easily be verified that Eq. (28) is in agreement with Eq. (25) up to the first-order terms in ω_0. Thus, whenever the frequency of band width variation is reasonably small by comparison with the mean band width, Eq. (28) provides an adequate approximation to the impulsive response of the network. A higher order approximation may be obtained, if necessary, by continuing the iteration beyond F_2.

An Extension of Wiener's Theory of Prediction*

LOTFI A. ZADEH AND JOHN R. RAGAZZINI
Department of Electrical Engineering, Columbia University, New York, New York
(Received November 14, 1949)

The theory of prediction described in this paper is essentially an extension of Wiener's theory. It differs from the latter in the following respects.

1. The signal (message) component of the given time series is assumed to consist of two parts, (a) a non-random function of time which is representable as a polynomial of degree not greater than a specified number n and about which no information other than n is available; and (b) a stationary random function of time which is described statistically by a given correlation function. (In Wiener's theory, the signal may not contain a non-random part except when such a part is a known function of time.)
2. The impulsive response of the predictor or, in other words, the weighting function used in the process of prediction is required to vanish outside of a specified time interval $0 \leq t \leq T$. (In Wiener's theory T is assumed to be infinite.)

The theory developed in this paper is applicable to a broader and more practical class of problems than that covered in Wiener's theory. As in Wiener's theory, the determination of the optimum predictor reduces to the solution of an integral equation which, however, is a modified form of the Wiener-Hopf equation. A simple method of solution of the equation is developed. This method can also be applied with advantage to the solution of the particular case considered by Wiener. The use of the theory is illustrated by several examples of practical interest.

I. INTRODUCTION

PREDICTION—in the broad sense of the term—consists essentially of estimating the values of some function of time on the basis of a time series which may or may not contain random errors. For instance, a typical problem in prediction is as follows. Given a time series $e_1(t)$ which is composed of a signal $s(t)$ and a random disturbance (noise) $N(t)$; provide an estimate of $s(t+\alpha)$, α being a positive constant, as a continuous function of time. More generally, the quantity to be estimated may be a functional of $s(t)$ such as ds/dt, $\int s dt$, etc. In forming such estimates the mathematical operations that may be employed are usually limited by practical considerations. Thus, in most cases the operator furnishing the estimate must be linear and fixed in addition to the obvious requirement of being physically realizable. The physical counterpart of such an operator is what is commonly known as a predictor or an estimator.

It is evident that a function of time cannot be predicted intelligently unless sufficient *a priori* information is available about both the function and the errors. The nature of such information, as well as the characteristics of the signal and noise, can assume a variety of forms. Of these the more common ones have been investigated in recent years with the result that for certain classes of time series it is now possible to design predictors which make optimum use of the *a priori* information concerning the signal and the noise. Thus, when the given time series is stationary and the correlation functions of the signal and noise are known, one can use Wiener's theory[1] to arrive at the specifications of the optimum predictor, that is, one minimizing the mean-square value of the prediction error. On the other hand, when, as is often the case in practice, the given time series is non-stationary, the available theories of prediction, notably Phillips and Weiss' theory,[2] do not lead to the best possible predictor except for a narrow class of time series. It is possible, however, to extend Wiener's theory in many different directions thereby making it applicable to a wider class of problems than is covered by either Wiener's or Phillips and Weiss' theories in their present forms. One such extension is discussed in this paper. A feature of the extension is that the signal (message) is assumed to consist of a stationary component superimposed on a non-random function of time which is known to be representable as a polynomial of degree less than or equal to a specified number n. It will also be shown that the general method developed for treating this problem can be applied with advantage to the solution of many cases of practical

* This work was performed in association with the Special Projects Department of The M. W. Kellogg Company for the Watson Laboratories of the Air Materiel Command.

[1] N. Wiener, "The extrapolation, interpolation, and smoothing of stationary time series," Report of the Services 19, Research Project DIC-6037, M.I.T. (February, 1942). Published in book form by John Wiley and Sons, Inc., New York (1949).

[2] R. S. Phillips and P. R. Weiss, "Theoretical calculation on best smoothing of position data for gunnery prediction," Report 532, Rad. Lab., M.I.T. (February, 1944).

interest as well as the particular case considered by Wiener.

II. FORMULATION

Consider a given time series $e_1(t)$ which is the sum of a function $s(t)$ (signal) and a stationary random disturbance $N(t)$. Let $s^*(t)$ be the quantity to be estimated and let $s^*(t)$ be related to $s(t)$ through a given linear operator $K(p)$, i.e.,

$$s^*(t)=K(p)s(t). \tag{1}$$

$K(p)$ may be thought of as the system function of an *ideal predictor*, i.e., a predictor capable of perfect prediction in the absence of noise. In many cases, particularly those involving actual prediction, the operator $K(p)$ is not physically realizable so that the process of estimation cannot be carried out exactly even in the absence of random disturbances.

Frequently it will be convenient to use a different, though equivalent representation of Eq. (1), i.e.,

$$s^*(t)=\int_{-\infty}^{\infty} k(\tau)s(t-\tau)d\tau, \tag{2}$$

where τ is the variable of integration and $k(t)$ represents the impulsive response of the ideal predictor. $K(p)$ shall be referred to as the ideal prediction operator. As a matter of convenience, the more common of the many possible forms which $K(p)$ and $k(t)$ can assume are given in Table I.

Like all theories of prediction, the theory to be described applies only to a special class of time series. The time series to be considered in the work which follows will be assumed to consist of a signal $s(t)$ and noise $N(t)$, with the signal being composed of a random component $M(t)$ superposed upon a non-random function of time $P(t)$, i.e.,

$$s(t)=M(t)+P(t). \tag{3}$$

The assumptions made concerning the characteristics of $P(t)$, $M(t)$, and $N(t)$, are as follows:

(a) $P(t)$ is assumed to be representable as a polynomial in t of degree not higher than a specified number n.

(b) $M(t)$ and $N(t)$ are stationary functions of time described respectively by their auto-correlation functions $\psi_M(\tau)$ and $\psi_N(\tau)$.

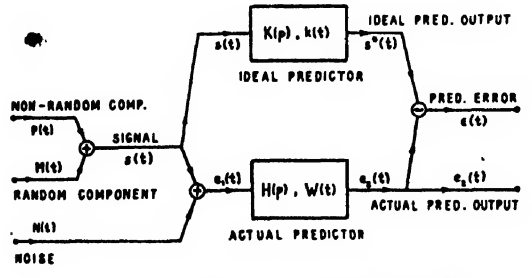


FIG. 1. Flow diagram of prediction process.

TABLE I. Common forms of the prediction operators $K(p)$ and $k(t)$.

Relation between $s^*(t)$ and $s(t)$	Significance of the quantity to be estimated	$K(p)$	$k(t)$
$s^*(t)=s(t)$	Present value of $s(t)$	1	$\delta(t)$
$s^*(t)=\dot{s}(t)$	Present value $\dot{s}(t)$	p	$\delta^{(1)}(t)$
$s^*(t)=\ddot{s}(t)$	Present value of $\dot{s}(t)$	p^2	$\delta^{(2)}(t)$
$s^*(t)=s(t-\alpha)$	Past or future value of $s(t)$ depending respectively on whether α is a positive or negative constant	$e^{-\alpha p}$	$\delta(t-\alpha)$

Note.—$\delta(t)$ denotes a unit impulse at $t=0$, and $\delta^{(\nu)}(t)$ stands for the νth derivative of $\delta(t)$ with respect to t (time).

(c) $M(t)$ and $N(t)$ have zero mean and are uncorrelated. This assumption is introduced only for the purpose of simplification and is not essential to the analysis. The condition expressed by (c) prevails in most practical cases.

Referring to Fig. 1, these inputs are shown being applied to the actual predictor whose system function is $H(p)$ and whose impulsive response is $W(t)$. The output of the predictor, $e_2(t)$, may be expressed in operational form

$$e_2(t)=H(p)e_1(t) \tag{4}$$

or, alternatively, in the form of a superposition integral

$$e_2(t)=\int_0^{\infty} W(\tau)e_1(t-\tau)d\tau. \tag{5}$$

An important characteristic of the actual predictor is the so-called prediction or estimation error ϵ, which is defined as the difference between the output of the predictor and the quantity to be estimated, $s^*(t)$. Equation-wise this is:

$$\epsilon=e_2(t)-s^*(t). \tag{6}$$

If there were no noise and if $K(p)$ were physically realizable there would be no prediction error and $H(p)$ would be identical with $K(p)$. This, of course, is the trivial case of the prediction problem. In what follows it will be assumed that either because of the presence of noise or physical unrealizability of $K(p)$, or both, $H(p)$ cannot be the same as $K(p)$.

The available *a priori* information about $s(t)$ and $N(t)$ is assumed to consist of n, $\psi_M(\tau)$, and $\psi_N(\tau)$. The problem is to specify the system function or the impulsive response of a predictor that would minimize in a certain sense the prediction error $\epsilon=e_2(t)-s^*(t)$. By analogy with Wiener's theory it will be postulated that the optimum predictor is the one in which: (a) the ensemble mean of ϵ is equal to zero (for all values of t), and (b) the ensemble variance of ϵ is a minimum. Denoting the ensemble average by the symbol $\langle\ \rangle_{Av}$, these conditions read:

$$\text{(a)}\quad \langle\epsilon\rangle_{Av}=0 \text{ or, equivalently, } \langle e_2(t)\rangle_{Av}\equiv\langle s^*(t)\rangle_{Av}, \tag{7}$$

$$\text{(b)}\quad \sigma^2=\langle\epsilon^2\rangle_{Av}=\text{minimum}, \tag{8}$$

where σ^2, the ensemble variance of ϵ, is equal to the mean-square value of the prediction error. In what follows, conditions (a) and (b) will be used as the basis for the determination of the optimum predictor.

III. DETERMINATION OF THE IMPULSIVE RESPONSE OF THE OPTIMUM PREDICTOR

It will be recalled that the output of a predictor may be expressed in the form of a superposition integral

$$e_2(t)=\int_0^\infty W(\tau)e_1(t-\tau)d\tau \tag{9}$$

where τ is a dummy variable and $W(t)$ represents the impulse response of the predictor. In practice it is usually found necessary to restrict the duration of sampling of the input time series to a finite constant T, meaning in other words that $W(t)$ must be zero outside the interval $0\leq t\leq T$. To place this property in evidence Eq. (9) will be written in the following form:

$$e_2(t)=\int_0^T W(\tau)e_1(t-\tau)d\tau. \tag{10}$$

In the limiting case where the duration of sampling is infinite ($T\to\infty$) Eq. (10) becomes identical with Eq. (9).

By hypothesis,

$$e_1(t)=P(t)+M(t)+N(t). \tag{11}$$

Substituting Eq. (11) into Eq. (10) and making use of the identity

$$P(t-\tau)\equiv P(t)-\tau\dot{P}(t)+\frac{\tau^2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\tau^n}{n!}P^{(n)}(t), \tag{12}$$

it is found that $e_2(t)$ may be expressed as:

$$e_2(t)=\mu_0P(t)-\mu_1\dot{P}(t)+\frac{\mu_2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\mu_n}{n!}P^{(n)}(t)+\int_0^T W(\tau)M(t-\tau)d\tau+\int_0^T W(\tau)N(t-\tau)d\tau, \tag{13}$$

where μ_0, μ_1, μ_2, etc., designate the moments of $W(t)$, i.e.,

$$\mu_\nu=\int_0^T \tau^\nu W(\tau)d\tau,\quad \nu=0, 1, 2, \cdots n. \tag{14}$$

Since $M(t)$ and $N(t)$ are stationary (with zero mean), it follows that the ensemble means of $e_2(t)$ and $s^*(t)$ depend only on the non-random component of the signal, i.e.,

$$\langle e_2(t)\rangle_{Av}=\int_0^T W(\tau)P(t-\tau)d\tau \tag{15}$$

or

$$\langle e_2(t)\rangle_{Av}=\mu_0P(t)-\mu_1\dot{P}(t)+\frac{\mu_2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\mu_n}{n!}P^{(n)}(t), \tag{16}$$

and

$$\langle s^*(t)\rangle_{Av}=\langle K(p)s(t)\rangle_{Av} \tag{17}$$

or

$$\langle s^*(t)\rangle_{Av}=K(p)P(t). \tag{18}$$

Comparing Eqs. (16) and (18), condition (a) is reduced to

$$K(p)P(t)\equiv\mu_0P(t)-\mu_1\dot{P}(t)+\frac{\mu_2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\mu_n}{n!}P^{(n)}(t). \tag{19}$$

Equation (19), being an identity, determines the values of μ_0, μ_1, $\cdots$, and μ_n. In other words, *the ideal prediction operator $K(p)$ determines through Eq. (19) the first $n+1$ moments of the impulsive response of the optimum predictor.*

As an illustration of the foregoing statement consider a case where the quantity to be estimated is the derivative of $s(t)$, i.e., $s^*(t)=\dot{s}(t)$. For this case Eq. (19) reduces to

$$\dot{P}(t)\equiv\mu_0P(t)-\mu_1\dot{P}(t)+\frac{\mu_2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\mu_n}{n!}P^{(n)}(t), \tag{20}$$

and a term by term comparison of the left-hand and right-hand sides of Eq. (20) yields:

$$\begin{aligned}\mu_0&=\int_0^T W(\tau)d\tau=0\\ \mu_1&=\int_0^T \tau W(\tau)d\tau=-1\\ \mu_2&=\int_0^T \tau^2W(\tau)d\tau=0\\ &\cdots\\ \mu_n&=\int_0^T \tau^nW(\tau)d\tau=0.\end{aligned} \tag{21}$$

These, therefore, are the $n+1$ constraints which the impulsive response of a derivative estimating network must satisfy.

As the second example consider a case where $K(p)s(t)=s(t-\alpha)$, α being a positive or negative constant. For

this case Eq. (19) reads

$$P(t-\alpha)\equiv\mu_0 P(t)-\mu_1\dot{P}(t)+\frac{\mu_2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\mu_n}{n!}P^{(n)}(t). \quad (22)$$

Rewriting $P(t-\alpha)$ as

$$P(t-\alpha)\equiv P(t)-\alpha\dot{P}(t)+\frac{\alpha^2}{2!}\ddot{P}(t)+\cdots+(-1)^n\frac{\alpha^n}{n!}P^{(n)}(t), \quad (23)$$

and making in Eq. (22) a term-by-term comparison of the coefficients of $P(t)$, $\dot{P}(t)$, etc., it is easily found that:

$$\begin{aligned}\mu_0&=\int_0^T W(\tau)d\tau=1\\ \mu_1&=\int_0^T \tau W(\tau)d\tau=\alpha\\ &\cdots\\ \mu_n&=\int_0^T \tau^n W(\tau)d\tau=\alpha^n,\end{aligned} \quad (24)$$

which thus represent the constraints imposed upon $W(t)$ in case the quantity to be estimated is $s(t-\alpha)$.

The problem that remains to be solved is that of minimizing σ^2. For this purpose it will be necessary to develop an explicit expression for σ^2 in terms of $W(t)$ and the auto-correlation functions of the signal and noise. Assuming that condition (a) is satisfied, it follows from inspection of Eqs. (6), (13), and (19) that the prediction error is given by the expression

$$\epsilon=\int_0^T W(\tau)[M(t-\tau)+N(t-\tau)]d\tau-K(p)M(t) \quad (25)$$

or equivalently

$$\epsilon=\int_0^T W(\tau)[M(t-\tau)+N(t-\tau)]d\tau-\int_{-\infty}^{\infty}k(\tau)M(t-\tau)d\tau \quad (26)$$

where $k(t)$ is the impulsive response of the ideal predictor. The mean-square value of ϵ may be written as

$$\sigma^2=\langle\epsilon^2\rangle_{Av}=\lim_{L\to\infty}\frac{1}{L}\int_0^L \epsilon^2 dt \quad (27)$$

or

$$\sigma^2=\lim_{L\to\infty}\frac{1}{L}\int_0^L dt\left\{\int_0^T W(\tau)[M(t-\tau)+N(t-\tau)]d\tau-\int_{\infty}^{\infty}k(\tau)M(t-\tau)d\tau\right\}^2. \quad (28)$$

A typical term of Eq. (28) such as

$$\lim_{L\to\infty}\frac{1}{L}\int_0^L dt\left[\int_0^T W(\tau)M(t-\tau)d\tau\right]^2 \quad (29)$$

is expressible in the form of a triple integral

$$\int_0^T\int_0^T d\tau_1 d\tau_2 W(\tau_1)W(\tau_2)\times\lim_{L\to\infty}\frac{1}{L}\int_0^L M(t-\tau_1)M(t-\tau_2)dt \quad (30)$$

which in view of the definition of the auto-correlation function of $M(t)$, i.e.,

$$\psi_M(\tau)=\lim_{L\to\infty}\frac{1}{L}\int_0^L M(t)M(t-\tau)dt \quad (31)$$

may be written as

$$\int_0^T\int_0^T W(\tau_1)W(\tau_2)\psi_M(\tau_1-\tau_2)d\tau_1 d\tau_2. \quad (32)$$

Proceeding similarly in the case of other terms, Eq. (28) reduces finally to the following expression:

$$\begin{aligned}\sigma^2=&\int_0^T\int_0^T W(\tau_1)W(\tau_2)[\psi_M(\tau_1-\tau_2)\\ &+\psi_N(\tau_1-\tau_2)]d\tau_1 d\tau_2\\ &-2\int_{\infty}^{\infty}\int_0^T W(\tau_1)k(\tau_2)\psi_M(\tau_1-\tau_2)d\tau_1 d\tau_2\\ &+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}k(\tau_1)k(\tau_2)\psi_M(\tau_1-\tau_2)d\tau_1 d\tau_2\end{aligned} \quad (33)$$

where, to recapitulate: τ_1, τ_2 = dummy variables; $W(t)$ = impulsive response of the predictor; $\psi_M(\tau)$ = auto-correlation function of $M(t)$ [$M(t)$ is the stationary part of the input signal]; $\psi_N(\tau)$ = auto-correlation function of $N(t)$ [$N(t)$ is the input noise]; $k(t)$ = impulsive response of the ideal predictor.

Returning to the problem of minimization of σ^2 it will be noted first that the last term in Eq. (33) is independent of $W(t)$ and hence, insofar as minimization of σ^2 is concerned, need not be considered. Second, it will be recalled that $W(t)$ is subject to the $n+1$ constraints

expressed by Eq. (14); therefore, the problem of minimizing σ^2 with respect to the class of $W(t)$'s satisfying Eq. (14) reduces essentially to an isoperimetric problem in the calculus of variations. Following the standard approach to such problems, one is led to minimizing the following expression:

$$I=\sigma^2-2\lambda_0\mu_0-2\lambda_1\mu_1-\cdots-2\lambda_n\mu_n \tag{34}$$

or, more explicitly

$$I=\int_0^T W(\tau_1)d\tau_1\Big\{\int_0^T W(\tau_2)[\psi_M(\tau_1-\tau_2)+\psi_N(\tau_1-\tau_2)]d\tau_2-2\int_{-\infty}^{\infty}k(\tau_2)\psi_M(\tau_1-\tau_2)d\tau_2-2\lambda_0-2\lambda_1\tau_1-\cdots-2\lambda_n\tau_1^n\Big\} \tag{35}$$

where the constants $\lambda_0, \lambda_1, \cdots, \lambda_n$, are the Lagrangian multipliers. Proceeding in the usual manner, that is, setting the variation of I equal to zero, it is easily found that I and hence σ^2 is a minimum provided $W(t)$ satisfies the following integral equation:

$$\int_0^T W(\tau)[\psi_M(t-\tau)+\psi_N(t-\tau)]d\tau=\lambda_0+\lambda_1 t+\cdots+\lambda_n t^n+\int_{-\infty}^{\infty}k(\tau)\psi_M(t-\tau)d\tau,\quad 0\leq t\leq T. \tag{36}$$

This equation together with the $n+1$ constraints expressed by Eq. (14) provides the basis for the determination of the optimum predictor. It will be observed that in the particular case where $n=0$, $T=\infty$, and $k(t)=\delta(t+\alpha)$ [$\delta(t)$ standing, as usual, for a unit impulse at $t=0$], Eq. (36) reduces to

$$\int_0^{\infty} W(\tau)[\psi_M(t-\tau)+\psi_N(t-\tau)]d\tau=\psi_M(t+\alpha),\quad t>0 \tag{37}$$

which is essentially the integral equation of Wiener's theory. On the other hand, in the special case where $M(t)\equiv 0$, Eq. (36) reduces to

$$\int_0^T W(\tau)\psi_N(t-\tau)d\tau=\lambda_0+\lambda_1 t+\cdots+\lambda_n t^n,\quad 0\leq t\leq T \tag{38}$$

which is the integral equation of Phillips and Weiss' theory. Thus, the integral equations of Wiener's, and Phillips and Weiss' theories are special cases of Eq. (36).

IV. SOLUTION OF THE INTEGRAL EQUATION

In the general case where $\psi_M(\tau)$ and $\psi_N(\tau)$ are prescribed but otherwise arbitrary auto-correlation functions, the complicated nature of the integral equation makes it appear that the solution of Eq. (36) is a formidable problem. In reality, the problem is not as difficult as it may seem, for by using a procedure to be described, the general case can be reduced to a special case which has a simple solution.

Preliminary to the discussion of this procedure it will be expedient to introduce the spectral densities of $M(t)$, $N(t)$, and $M(t)+N(t)$. Denoting these by $S_M(\omega^2)$, $S_N(\omega^2)$, and $S(\omega^2)$, respectively, and recalling that the spectral density of a function is the inverse Fourier transform of its auto-correlation function,[1] it follows that

$$S_M(\omega^2)=\int_{-\infty}^{\infty}\psi_M(\tau)e^{-i\omega\tau}d\tau \tag{39}$$

$$S_N(\omega^2)=\int_{-\infty}^{\infty}\psi_N(\tau)e^{-i\omega\tau}d\tau \tag{40}$$

and

$$S(\omega^2)=S_M(\omega^2)+S_N(\omega^2). \tag{41}$$

Now the spectral density function $S(\omega^2)$ may be factored[1] into the product of two conjugate factors

$$S(\omega^2)=G(j\omega)\cdot G(-j\omega) \tag{42}$$

such that both $G(j\omega)$ and $1/G(j\omega)$ are analytic in the right half of the $j\omega$-plane. Usually $S(\omega^2)$ is assumed to be a rational function of ω^2 of the form

$$S(\omega^2)=[A(\omega^2)/B(\omega^2)], \tag{43}$$

where $A(\omega^2)$ and $B(\omega^2)$ are polynomials in ω^2. For such cases the process of factorization is quite straightforward as can be seen from the following examples:

(a) $S(\omega^2)=\omega^2$; $\quad G(j\omega)=j\omega$.

(b) $S(\omega^2)=\dfrac{1}{\omega^2+\omega_0^2}$; $\quad G(j\omega)=\dfrac{1}{j\omega+\omega_0}$.

(c) $S(\omega^2)=\dfrac{\omega^2+a^2}{\omega^4+b^2\omega^2+c^4}$;

$$G(j\omega)=\frac{j\omega+a}{(j\omega)^2+j\omega(b^2+2c^2)^{\frac{1}{2}}+c^2}.$$

To summarize, a rational spectral density function may be written as

$$S(\omega^2)=|G(p)|^2_{p=j\omega} \tag{44}$$

where $G(p)$ is of the form:

$$G(p)=\frac{Q(p)}{R(p)}=\frac{a_0+a_1p+\cdots+a_mp^m}{b_0+b_1p+\cdots+b_lp^l}, \tag{45}$$

and the polynomials $Q(p)$ and $R(p)$ do not have any zeros in the right half of the p-plane.

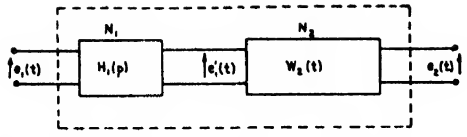


FIG. 2. Division of the predictor into the component networks N_1 and N_2.

The outline of the procedure used[3] for the solution of Eq. (36) can be best explained with reference to Fig. 2. The predictor N is assumed to be composed of two networks N_1 and N_2. The function of N_1 is to suitably modify some of the characteristics of the input time series $e_1(t)$, while that of N_2 is to provide the desired prediction through operating on the time series $e_1'(t)$, which is the output of N_1. It will be seen later that it is possible to choose N_1 in such a manner that the determination of the impulsive response of N_2 becomes an easily solvable problem. Then, once $W_2(t)$ (the impulsive response of N_2) is determined, the impulsive response of N, $W(t)$, can easily be found from the relation

$$W(t)=H_1(p)W_2(t), \tag{46}$$

where $H_1(p)$ is the system function of N_1. The choice of $H_1(p)$ and the problem of determination of $W_2(t)$ are discussed in the sequel.

It is evident that the problem of determination of $W_2(t)$ is similar to that of the determination of $W(t)$, except that the characteristics of the input time series are different for the two problems. An inspection of the integral equation (36) shows that it can be solved rather easily when the input to the predictor consists of a polynomial in t and a stationary component whose spectral density is a polynomial in ω^2. Therefore, in order to make the determination of $W_2(t)$ a simple problem, it is necessary to provide N_2 with an input which has this property. It is not difficult to verify that such a condition will obtain if, and only if, the system function of N_1 is chosen to be

$$H_1(p)=R(p), \tag{47}$$

where $R(p)$ is the denominator of $G(p)$ [cf. Eq. (45)]. With this choice of $H_1(p)$ the input to N_2 will consist of a polynomial in t of the same degree[4] as $P(t)$, and a stationary component $M'(t)+N'(t)$ whose spectral density is

$$S'(\omega^2)=|H_1(j\omega)|^2S(\omega^2) \tag{48}$$

or, in view of Eqs. (43), (44), and (47),

$$S'(\omega^2)=A(\omega^2), \tag{49}$$

where $A(\omega^2)$ is the numerator of $S(\omega^2)$. It will be noted that $A(\omega^2)$ is a polynomial of the form

$$A(\omega^2)=\gamma_0+\gamma_1\omega^2+\cdots+\gamma_m\omega^{2m}, \tag{50}$$

and correspondingly the auto-correlation function of $M'(t)+N'(t)$ is

$$\psi_{M'}(\tau)+\psi_{N'}(\tau)=\gamma_0\delta(\tau)-\gamma_1\delta^{(2)}(\tau)+\cdots+(-1)^m\gamma_m\delta^{(2m)}(\tau), \tag{51}$$

where $\delta^{(\nu)}(\tau)$ represents the impulse function of νth order [i.e., the νth derivative of the unit impulse function $\delta(\tau)$].

In addition to $\psi_{M'}(\tau)+\psi_{N'}(\tau)$, a number of other quantities associated with the input to N_2 enter the integral equation satisfied by $W_2(t)$. The significance of each of these quantities, as well as their expressions, are as follows:

(a) $S_{M'}(\omega^2)=$ spectral density of $M'(t)$
$$=S_M(\omega^2)|R(j\omega)|^2. \tag{52}$$

(b) $\psi_{M'}(\tau)=$ auto-correlation function of $M'(t)$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}S_M(\omega^2)|R(j\omega)|^2e^{j\omega\tau}d\omega. \tag{53}$$

(c) $k'(t)=$ ideal impulsive response for N_2
$$=[1/R(p)]k(t). \tag{54}$$

In terms of these quantities the integral equation satisfied by $W_2(t)$ reads:

$$\int_0^{\infty}W_2(\tau)[\psi_{M'}(t-\tau)+\psi_{N'}(t-\tau)]d\tau=\lambda_0'+\lambda_1't+\cdots+\lambda_n't^n+\int_{-\infty}^{\infty}k'(\tau)\psi_{M'}(t-\tau)d\tau, \quad t\geq 0. \tag{55}$$

It will be noticed that in the case of $W_2(t)$ the upper limit of the integral is infinity, while in the case of $W(t)$ [cf. Eq. (36)] it is T. The explanation for this difference is that $W_2(t)$ need not vanish for $t>T$, even though $W(t)$ is required to do so. Thus in general, $W_2(t)$ will be piecewise analytic in the interval $0<t<\infty$ as is illustrated in Fig. 3. Denoting the parts of $W_2(t)$ extending over the intervals $0\leq t\leq T$ and $T<t<\infty$ by $U(t)$ and $V(t)$, respectively, the relation connecting $W(t)$ and $W_2(t)$ [cf. Eq. (46)] may be rewritten in the following form:

$$W(t)=R(p)U(t) \tag{56a}$$

and

$$0=R(p)V(t). \tag{56b}$$

These relations show that $W(t)$ is completely determined by the part of $W_2(t)$ which extends over the interval $0\leq t\leq T$; the form of $W_2(t)$ outside this interval is irrelevant to the determination of $W(t)$.

[3] The appendix of a report by Bode, Blackman, and Shannon, "Data smoothing and prediction in fire-control systems," Research and Development Board, Washington, D. C. (August, 1948), contains a brief exposition of a method which is similar in certain respects to the method described here.

[4] It is tacitly assumed that $R(p)$ does not have a zero at the origin or, in other words, that $S(\omega^2)$ does not have a pole at zero frequency.

Returning to the integral equation (55), it will be noted that the range of integration $0\leq\tau<\infty$ may be divided into two parts, $0\leq\tau\leq T$ and $T<\tau<\infty$, involving $U(t)$ and $V(t)$, respectively. Since $V(t)$ is determined by Eq. (56b) to within a finite number of constants, the integral equation in question degenerates into an integral equation involving only $U(t)$:

$$\int_0^T U(\tau)[\psi_M'(t-\tau)+\psi_N'(t-\tau)]d\tau=\lambda_0'+\lambda_1't+\cdots+\lambda_n't^n+\int_{-\infty}^{\infty}k'(\tau)\psi_M'(t-\tau)d\tau,\quad 0\leq t\leq T. \quad (57)$$

Upon substitution of Eqs. (52), (53), and (54), and performing minor simplifications, Eq. (57) reads

$$\int_0^T U(\tau)[\gamma_0\delta(t-\tau)-\gamma_1\delta^{(2)}(t-\tau)+\cdots+(-1)^m\gamma_m\delta^{(2m)}(t-\tau)]d\tau=\lambda_0'+\lambda_1't+\cdots+\lambda_n't^n+\frac{1}{2\pi}\int_{-\infty}^{\infty}S_M(\omega^2)K(j\omega)R(-j\omega)e^{j\omega t}d\omega. \quad (58)$$

Making use of the identity

$$\int_0^T U(\tau)\delta^{(2\nu)}(t-\tau)d\tau\equiv p^{2\nu}U(t), \quad (59)$$

Equation (58) may be rewritten as

$$[\gamma_0-\gamma_1p^2+\cdots+(-1)^m\gamma_mp^{2m}]U(t)=\lambda_0'+\lambda_1't+\cdots+\lambda_n't^n+\frac{1}{2\pi}\int_{-\infty}^{\infty}S_M(\omega^2)K(j\omega)R(-j\omega)e^{j\omega t}d\omega. \quad (60)$$

Since in this equation the left-hand side operator is simply $A(-p^2)$ [cf. Eq. (50)], the integral equation (57) finally reduces to the following differential equation:

$$A(-p^2)U(t)=\lambda_0'+\lambda_1't+\cdots+\lambda_n't^n+\frac{1}{2\pi}\int_{-\infty}^{\infty}S_M(\omega^2)K(j\omega)R(-j\omega)e^{j\omega t}d\omega. \quad (61)$$

The general solution of this equation is of the form:

$$U(t)=A_0'+A_1't+\cdots+A_n't^n+B_1'\exp(\alpha_1t)+B_2'\exp(\alpha_2t)+\cdots+B_{2m}'\exp(\alpha_{2m}t)+\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{A(\omega^2)}K(j\omega)R(-j\omega)e^{j\omega t}d\omega,\quad 0\leq t\leq T, \quad (62)$$

where A_0', A_1', $\cdots$, A_n' and B_1', B_2', $\cdots$, B_{2m}' are as yet undetermined constants, and α_1, α_2, $\cdots$, α_{2m} are the roots of the characteristic equation

$$A(-p^2)=0. \quad (63)$$

In brief, Eq. (62) provides an explicit expression for $U(t)$ involving $2m+n+1$ undetermined constants. Availability of such an expression reduces the problem of determination of $W(t)$ to a relatively routine matter which is discussed in the following section.

V. DERIVATION OF AN EXPLICIT EXPRESSION FOR $W(t)$

Recalling that $W(t)$ is related to $U(t)$ through the operational relation

$$W(t)=R(p)U(t), \quad (56a)$$

and substituting $U(t)$ as given by Eq. (62) into Eq. (56a), it is readily found that in the most general case $W(t)$ is given by the following expression:

$$W(t)=[u(t)-u(t-T)]\Big\{A_0+A_1t+\cdots+A_nt^n+B_1\exp(\alpha_1t)+\cdots+B_{2m}\exp(\alpha_{2m}t)+\frac{1}{2\pi}R(p)\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{A(\omega^2)}K(j\omega)R(-j\omega)e^{j\omega t}d\omega\Big\}+C_1\delta(t)+\cdots+C_{l-m}\delta^{(l-m-1)}(t)+D_1\delta(t-T)+\cdots+D_{l-m}\delta^{(l-m-1)}(t-T) \quad (64)$$

where the A's, B's, C's, and D's are as yet undetermined constants, and the unit step functions $u(t)$ and $u(t-T)$ are used simply to indicate that $W(t)$ is zero outside the interval $0\leq t\leq T$. The impulse functions contained in the expression for $W(t)$ arise from operation by $R(p)$ on the discontinuities of $U(t)$ at $t=0$ and $t=T$. It will be observed that the order of these impulse functions does not exceed $l-m-1$, which is one-half the order of the zero of $S(\omega^2)$ at infinity minus one. This is due to the fact that the first $m-1$ derivatives of $U(t)$ vanish at $t=0$ and $t=T$. It is not difficult to verify that if this would not have been the case, the mean-square error at the output of N_2 would be infinite.

Having obtained the general expression for $W(t)$ in the form of Eq. (64), there remains the problem of determination of the $2l+n+1$ unknown constants. These can be found in the following manner:

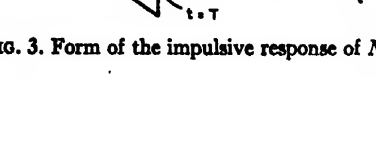


FIG. 3. Form of the impulsive response of N_2.

1. Substituting $W(t)$ as given by Eq. (64) into the integral equation (36) and requiring that the equation be satisfied identically, leads to $2l$ linear homogeneous equations in the A's, B's, C's, and D's.

2. Substituting $W(t)$ as given by Eq. (64) into the $n+1$ moment equations

$$\int_0^T \tau^\nu W(\tau)d\tau=\mu_\nu, \quad \nu=0, 1, 2, \cdots, n \qquad (65)$$

yields other $n+1$ linear equations. These $n+1$ equations, together with the $2l$ equations obtained in (1), provide a system of $2l+n+1$ linear equations in the unknown constants. Solution of this system gives the values of the A's, B's, C's, and D's and thus completes the process of determination of $W(t)$.

It should be remarked that in some cases it is advantageous to deal with the system function $H(p)$ of the predictor, rather than with its impulsive response $W(t)$. In such cases one can use a transformed form of the integral equation (36) which is as follows:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega)S(\omega^2)e^{j\omega t}d\omega=\lambda_0+\lambda_1 t+\cdots+\lambda_n t^n$$
$$+\frac{1}{2\pi}\int_{-\infty}^{\infty} S_M(\omega^2)K(j\omega)R(-j\omega)e^{j\omega t}d\omega. \qquad (66)$$

Using Eq. (64), the solution of this equation may be written directly as

$$H(p)=\int_0^T (A_0+A_1 t+\cdots+A_n t^n)e^{-pt}dt$$
$$+\frac{B_1}{p+\alpha_1}+\frac{B_2}{p+\alpha_2}+\cdots+\frac{B_{2m}}{p+\alpha_{2m}}$$
$$-\left\{\frac{B_1\exp(\alpha_1 T)}{p+\alpha_1}+\frac{B_2\exp(\alpha_2 T)}{p+\alpha_2}+\cdots\right.$$
$$\left.+\frac{B_{2m}\exp(\alpha_{2m}T)}{p+\alpha_{2m}}\right\}e^{-pT}+\frac{1}{2\pi}R(p)\int_0^T dte^{-pt}$$
$$\times\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{A(\omega^2)}K(j\omega)R(-j\omega)e^{j\omega t}d\omega$$
$$+C_1+C_2p+\cdots+C_{l-m}p^{l-m-1}$$
$$+(D_1+\cdots+D_{l-m}p^{l-m-1})e^{-pT}. \qquad (67)$$

The undetermined constants involved in this expression are found in the same manner as in the case of $W(t)$, that is, $H(p)$ as given by Eq. (67) is substituted into the integral equation (66) and the resulting expression is treated as an identity. The $2l$ linear relations between $A_0, A_1, A_2, \cdots$, etc., which are obtained in this manner are adjoined to the $n+1$ relations resulting from Eq. (65); then the system of linear equations in the unknown constants is solved for $A_0, A_1, A_2, \cdots$, etc.

In order to facilitate application of the techniques described in the preceding sections, a summary of the procedure for the determination of $W(t)$ (or $H(p)$) is given in Section VI. Furthermore, actual use of the procedure is illustrated by a few practical examples at the end of the section.

VI. SUMMARY OF THE PROCEDURE FOR DETERMINATION OF $W(t)$ AND $H(p)$

The complete expressions for $W(t)$ (the impulsive response of the optimum predictor) and $H(p)$ (the system function of the optimum predictor) are given by Eqs. (64) and (67). In order to avoid the necessity for reference to preceding sections, the meaning of all symbols appearing in these equations is given:

$u(t)$ = unit step function.
T = duration of sampling (settling time).
$A_0, A_1, \cdots, A_n, B_1, B_2, \cdots, B_{2m}, C_1, C_2, \cdots, C_{l-m}, D_1, D_2, \cdots, D_{l-m}$ = undetermined constants.
n = degree of the polynomial component of the input signal.
$S_M(\omega^2)$ = spectral density of $M(t)$[$M(t)$ is the stationary part of the input signal].
$S(\omega^2)$ = spectral density of $M(t)+N(t)$[$N(t)$ is the input noise].
$A(\omega^2)$ = numerator of $S(\omega^2)$.
$B(\omega^2)$ = denominator of $S(\omega^2)$.
$Q(j\omega)$ = a factor of $A(\omega^2)$ containing all the zeros in right half of the $j\omega$-plane.
$R(j\omega)$ = a factor of $B(\omega^2)$ containing all the zeros in the left half of the $j\omega$-plane.
$2l$ = degree of $B(\omega^2)$.
$2m$ = degree of $A(\omega^2)$.
$\alpha_1, \alpha_2, \cdots, \alpha_{2m}$ = roots of the characteristic equation $A(-p^2)=0$.
$\delta(t)$ = unit impulse function.
$\delta^{(\nu)}(t)$ = νth derivative of $\delta(t)$.

The undetermined constants occurring in the expression for $W(t)$ [and $H(p)$] can be found in the following manner.

1. $W(t)$ as given by Eq. (64) is substituted into the integral equation (36) and the resulting expression is treated as an identity. This gives $2l$ homogeneous linear equations in the unknown constants. Same equations can be obtained by substituting $H(p)$, as given by Eq. (67), into the integral equation (66).

2. $W(t)$ as given by Eq. (64) is substituted into the $n+1$ constraint equations

$$\int_0^T \tau^\nu W(\tau)d\tau=\mu_\nu, \quad \nu=0, 1, \cdots, n. \qquad (14)$$

where the μ_ν are determined by the choice of the prediction operator $K(p)$ [cf. Eq. (19)]. The resulting $n+1$ linear equations in the unknown constants are adjoined to the $2l$ equations obtained from (1). The set of $2l+n+1$ linear equations thus obtained is solved for the undetermined constants $A_0, A_1, \cdots, D_{l-m}$. This concludes the process of determining $W(t)$ [or $H(p)$].

VII. ILLUSTRATIVE EXAMPLES

Example 1. Wiener's Theory

Wiener's theory is, in the main, a study of the particular case in which $P(t)\equiv 0$, $T=\infty$ and $K(p)=e^{-ap}$. For this case Eq. (64) gives

$$W(t)=u(t)\left\{\frac{1}{2\pi}R(p)\times\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{A(\omega^2)}K(j\omega)R(-j\omega)e^{j\omega(t-\alpha)}d\omega+B_1\exp(\alpha_1 t)+\cdots+B_{2m}\exp(\alpha_{2m}t)\right\}. \quad (68)$$

The exponential terms appearing in Eq. (68) may be made to vanish through a slight rearrangement of the factors in the first term of Eq. (68). The resulting expression for $W(t)$ is the same as that obtained by using Wiener's theory, namely,

$$W(t)=u(t)\frac{1}{2\pi}\frac{R(p)}{Q(p)}\times\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{Q(-j\omega)}R(-j\omega)e^{j\omega(t-\alpha)}d\omega. \quad (69)$$

The rearrangement amounts, essentially, to choosing a particular solution of Eq. (61) which differs from the one chosen before by the exponential terms of Eq. (68). The same result may be achieved directly by choosing $H_1(p)$ [cf. Eq. (47)] as

$$H_1(p)=R(p)/Q(p).$$

With this choice of $H_1(p)$ [in place of the one expressed by Eq. (47)] the various quantities entering Eq. (55) become:

$$S'(\omega^2)=1,$$
$$\psi_M'(\tau)+\psi_N'(\tau)=\delta(\tau),$$
$$S_M'(\omega^2)=S_M(\omega^2)|R(j\omega)/Q(j\omega)|^2,$$
$$K'(j\omega)=K(j\omega)Q(j\omega)/R(j\omega),$$

and hence the integral equation (55) reduces to

$$W_2(t)=u(t)\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{S_M(\omega^2)}{Q(-j\omega)}K(j\omega)R(-j\omega)e^{j\omega t}d\omega; \quad (55a)$$

Eq. (69) then follows immediately from the relation connecting $W(t)$ and $W_2(t)$ [cf. Eq. (46)].

Example 2

The assumptions made here are as follows:

1. $M(t)\equiv 0$.
2. $n=1$.
3. $\psi_N(\tau)=e^{-a|\tau|}$; $S_N(\omega^2)=\dfrac{2a}{\omega^2+a^2}$.

The choice of the prediction operator is left open.

Solution

For this case $A(-p^2)=2a$, and hence $\alpha_1=\cdots=\alpha_{2m}=0$. Also, $l=1$, $M\equiv 0$; hence Eq. (64) gives

$$W(t)=A_0+A_1t+C_1\delta(t)+D_1\delta(t-T), \quad (70)$$

and

$$H(p)=\frac{A_0}{p}(1-e^{-pT})+\frac{A_1}{p}\left(\frac{1}{p}-\frac{e^{-pT}}{p}+Te^{-pT}\right)+C_1+D_1e^{-pT}. \quad (71)$$

Substituting $H(p)$ as given by Eq. (71) into the integral equation

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{2a}{a^2-p^2}H(p)e^{pt}dp=\lambda_0+\lambda_1t,\quad 0\le t\le T, \quad (72)$$

and requiring that this equation be satisfied identically, leads to the following relations:

$$aA_0-A_1-a^2C_1=0, \quad (73)$$

$$aA_0+(aT+1)A_1-a^2D_1=0. \quad (74)$$

Furthermore, substituting $W(t)$ as given by Eq. (70) into the constraint equations

$$\int_0^T W(\tau)d\tau=\mu_0 \quad (75)$$

and

$$\int_0^T \tau W(\tau)d\tau=\mu_1, \quad (76)$$

yields

$$A_0T+A_1(T^2/2)+C_1+D_1=\mu_0, \quad (77)$$

and

$$A_0(T^2/2)+A_1(T^3/3)+D_1T=\mu_1. \quad (78)$$

The unknown constants A_0, A_1, C_1, and D_1 can be readily found from the solution of Eqs. (73), (74), (77), and (78). Thus,

$$A_0=\mu_0\frac{4a(a^2T^2+3aT+3)}{(a^2T^2+6aT+12)(aT+2)}-\mu_1\frac{6a^2}{a^2T^2+6aT+12}, \quad (79)$$

$$A_1=-\mu_0\frac{6a^2}{a^2T^2+6aT+12}+\mu_1\frac{12a^2}{T(a^2T^2+6aT+12)}, \quad (80)$$

$$C_1=\mu_0\frac{2(2a^2T^2+9aT+12)}{(a^2T^2+6aT+12)(aT+2)}-\mu_1\frac{6(aT+2)}{T(a^2T^2+6aT+12)}, \quad (81)$$

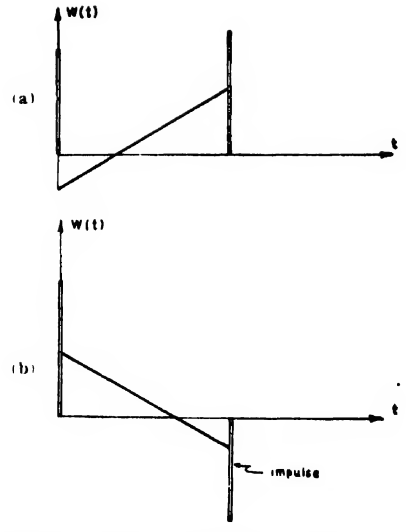


FIG. 4. Shapes of optimum weighting functions. (a) For the best value of $s(t)$. (b) For the best value of $s(t-\alpha)$ (α-negative).

and

$$D_1=-\mu_0\frac{2aT(aT+3)}{(a^2T^2+6aT+12)(aT+2)}+\mu_1\frac{6(aT+2)}{T(a^2T^2+6aT+12)}. \quad (82)$$

Whenever the stationary part of the input signal is zero [i.e., $M(t)\equiv 0$], the mean-square value of the prediction error assumes the simple form

$$\sigma^2=\mu_0\lambda_0+\mu_1\lambda_1+\cdots+\mu_n\lambda_n, \quad (83)$$

which may be readily established by substituting Eq. (36) into the general expression for σ^2 [cf. Eq. (33)]. For the particular case under consideration Eq. (83) gives

$$\sigma^2=\mu_0^2\frac{8(a^2T^2+3aT+3)}{(a^2T^2+6aT+12)(aT+2)}+\mu_1^2\frac{24a}{T(a^2T^2+6aT+12)}-\mu_0\mu_1\frac{24a}{(a^2T^2+6aT+12)}. \quad (84)$$

The expressions given above are valid for any choice of the prediction operator. For the particular case in which the predictor is called upon to furnish the best possible estimate of the present value of $\dot{s}(t)$, the values of μ_0 and μ_1 are, respectively,

$$\mu_0=0 \quad [\text{cf. Eq. (21)}] \quad \mu_1=-1. \quad (85)$$

On the other hand, in the case of the estimation of $s(t+\alpha)$ (i.e., the value of $s(t)$ α-seconds in the future)

$$\mu_0=1 \quad [\text{cf. Eq. (24)}] \quad \mu_1=-\alpha. \quad (86)$$

The shapes of $W(t)$ for these two particular cases are illustrated in Fig. 4.

Example 3

The case to be considered here is the same as that treated in Example 2, except that the auto-correlation function of $N(t)$ is assumed to be of the form

$$\psi_N(\tau)=e^{-a|\tau|}\cos\omega_0\tau, \quad (87)$$

with the associated spectral density function being

$$S_N(\omega^2)=\frac{2a(a^2+\omega_0^2+\omega^2)}{\omega^4+2(a^2-\omega_0^2)\omega^2+(a^2+\omega_0^2)^2}. \quad (88)$$

This form of spectral density function is of considerable practical importance since it provides a reasonably good approximation to many of the actual spectra encountered in practice.

Solution

By Eq. (64), the weighting function for this case is of the form

$$W(t)=A_0+A_1t+B_1e^{bt}+B_2e^{-bt}+C_1\delta(t)+D_1\delta(t-T), \quad 0\le t\le T, \quad (89)$$

where $b=(a^2+\omega_0^2)^{\frac{1}{2}}$. Substituting $W(t)$ as given by Eq. (89) into Eq. (36), and requiring that Eq. (36) be satisfied by $W(t)$ establishes four linear algebraic equations between the six constants A_0, A_1, B_1, B_2, C_1, D_1. These are:

$$\begin{aligned}
&-2ab^2A_0+2(a^2-\omega_0^2)A_1-b^3B_1+b^3B_2+C_1=0,\\
&2\omega_0^2b^2A_0-4a\omega_0^2A_1+b^3(b-a)B_1+b^3(b+a)B_2=0,\\
&-2ab^2A_0+2(\omega_0^2-a^2-aTb^2)A_1+\frac{e^{bT}}{2b}B_1-\frac{e^{-bT}}{2b}B_2+D_1=0,\\
&2\omega_0^2b^2A_0+2\omega_0^2(2a+b^2T)A_1+b^3(b+a)e^{bT}B_1+b^3(b-a)e^{-bT}B_2=0.
\end{aligned} \quad (90)$$

The constraints imposed by the prediction operator $K(p)$ are given by Eq. (14); they are:

$$\int_0^T W(\tau)d\tau=\mu_0, \quad (75)$$

and

$$\int_0^T \tau W(\tau)d\tau=\mu_1. \quad (76)$$

The requirement that $W(t)$ must satisfy Eqs. (75) and (76) leads to two additional linear equations. These are:

$$2bTA_0+bT^2A_1+2(e^{bT}-1)B_1-2(e^{-bT}-1)B_2 +2bC_1+2bD_1=\mu_0,$$

and

$$3b^2T^2A_0+2b^2T^3A_1+b[e^{bT}(bT-1)+1]B_1 -6[e^{-bT}(bT+1)-1]B_2+6b^2D_1=\mu_1. \quad (91)$$

The coefficients A_0, A_1, B_1, B_2, C_1, D_1 of Eq. (89) are the solutions of the six linear equations (90) and (91).

The expression for the mean-square value of the prediction error can be conveniently expressed in terms of A_0 and A_1. Thus, making use of Eq. (83) it is readily found that

$$\sigma^2=(2a/b^2)(\mu_0A_0+\mu_1A_1). \quad (92)$$

This completes essentially the solution of the problem.

ACKNOWLEDGMENTS

The authors wish to thank Mr. H. T. Marcy of The M. W. Kellogg Company and Mr. H. Sherman of the Watson Laboratories, AMC, for their encouragement and support of this work.

Potential Flow into Circumferential Openings in Drain Tubes*

DON KIRKHAM

Iowa State College Agricultural Experiment Station, Department of Agronomy and Department of Physics, Ames, Iowa

(Received December 22, 1949)

A theoretical analysis of the effect of the spaces between drain tube units as used in the artificial drainage of soil is given. The problem is one of potential flow; therefore, the results are applicable to heat flow, etc. The basic problem solved is that for axially symmetric flow from an external cylindrical boundary at constant potential to a series of equal, equally spaced openings at a lower potential, all located axially on, and comprising a part of, the otherwise impervious drain tube. The radii of the open sections and impermeable sections of the drain tube are equal. The basic problem is extended to obtain the solution to the practical problem—the seepage of ground water into drain tubes beneath a horizontal water table. The exact solution of the basic problem is not suitable for numerical work. Accordingly, approximate solutions of specified uncertainty are derived and are utilized for tabulation of numerical results. As an example, the analysis shows, in the case of 6 in. diameter drain tubes having 1 ft. long impermeable sections and buried 4 ft. deep in uniformly permeable soil, that increasing the openings from $\frac{1}{32}$ in. width to $\frac{1}{4}$ in. width will increase the flow 36 percent; while embedding the tubes in gravel, to make the $\frac{1}{32}$ in. openings of effectively infinite width, will increase the flow 180 percent.

IN the drainage of soil, for agricultural use, for foundations, roads, or for dams, excess ground water is generally removed by drain tubes installed two or more feet below the soil surface. These tubes consist, in most cases, of 1 ft. long sections of impervious pipe, either tile or cement, fitted together axially, except for a space left between individual units, to permit the water to enter. In practice the opening between the units is usually small. It may be less than 1/64 in. or, as much as $\frac{1}{4}$ in. But if the tubes are embedded in gravel, as is sometimes the case, the open space between the pipe units becomes effectively infinite. This is a consequence of the negligible loss of head which results when water seeps through gravel as compared with water seeping through soil. In this paper the effect of width of opening between the pipe units on drainage rate will be analyzed. As the problem is one of potential theory, the results will be applicable to other physical problems as flow of heat, electricity, etc.[1]

In the actual drainage problem the drain tubes are installed in a nearly horizontal position in the soil. Therefore if the soil is water-saturated to the surface of the ground, the soil surface is an equipotential plane, and the flow problem may be solved by the method of images.

AXIAL FLOW

Before applying the method of images it is convenient to solve a simpler problem: that of axially symmetrical flow for the system indicated in Fig. 1. Here, because of symmetry, only a section $ABCD$ need be considered. This section is bounded by inner and outer cylindrical surfaces at $r=a$ and $r=b$; by a plane $z=0$, midway between an open space of width $2c$; and a plane $z=s$, $2s$ being the distance between centers of openings.

Taking ϕ as the potential function, V_a as the constant potential over the openings of the drain tube, and V_b the constant potential over the outer cylinder, the boundary conditions are—I: $\phi=V_a$, $r=a$, $0\leq z\leq c$; II: $\partial\phi/\partial r=0$, $r=a$, $c<z\leq s$; III: $\partial\phi/\partial z=0$, $z=s$; IV: $\phi=V_b$, $r=b$; V: $\partial\phi/\partial z=0$, $z=0$.

The exact solution of $\nabla^2\phi$ which satisfies the above conditions can be formulated. This exact solution, is not, however, tractable in calculations. Therefore, a simpler, approximate solution is sought. Simplification is looked for through substitution of approximately correct,

* Journal Paper No. 1728 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project No. 998.

[1] M. Muskat, *The Flow of Homogeneous Fluids through Porous Media* (McGraw-Hill Book Company, Inc., New York, 1937, or J. W. Edwards, Inc., Ann Arbor, Michigan, 1946), p. 140.

homogeneous boundary conditions at $r=a$, for the accurate, mixed boundary conditions, I and II. Homogeneous boundary conditions at $r=a$ can be realized in two ways: (1) by using the exact condition, $\phi=V_a$ for $0<z<c$ and finding by trial a ϕ for $c<z<s$ such that over $c<z<s$, $\partial\phi/\partial r=0$; or (2), by using the exact condition

$$\partial\phi/\partial r=0, \quad c<z<s, \quad r=a \tag{1}$$

and finding by trial a $\partial\phi/\partial r$ such that

$$\partial\phi/\partial r=f(z, r, b, s, c)=f(z), \quad 0\leq z<c, \quad r=a, \tag{2}$$

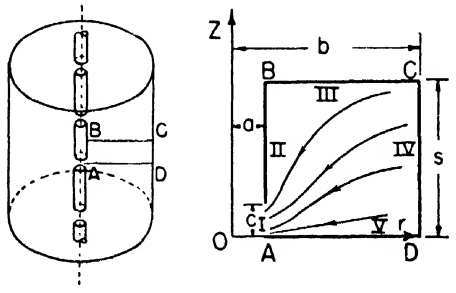


FIG. 1. Case of axially symmetric flow.

the ϕ in Eq. (2) satisfying simultaneously the relation

$$\phi=V_a, \quad 0\leq z<c, \quad r=a. \tag{3}$$

Because in practice one has $c\ll s$, with the consequence that $\partial\phi/\partial r$ is known exactly $(=0)$ over most of the boundary, it appears better to use the second method. Accordingly we attempt to satisfy Eqs. (1), (2) and (3).

As will appear, Eqs. (1) and (2) can, with the aid of Fourier series, be satisfied for an infinite number of values of $f(z)$. To satisfy, in addition, Eq. (3), however, there exists only one suitable value of $f(z)$, and to determine this value, by trial with accuracy, considerable labor would be required. A form of $f(z)$ which satisfies Eq. (3) for practical purposes can be determined by considering the two-dimensional case[2] into which Fig. 1 goes over when $a\to\infty$. Here with $b>s$, $s\gg c$, y replacing r, x replacing z, $z'=x'+iy'$, $i=(-1)^{\frac{1}{2}}$; one finds,[3] taking w for the complex potential,

$$\frac{dw}{dz'}=\frac{A\cos\pi(x'+iy')/2s}{[\sin^2\pi c/2s-\sin^2\pi(x'+iy')/2s]^{\frac{1}{2}}},$$

where A is a constant. Now, setting $y=0$, and observing that the x-component of dw/dz' over the opening is zero, there results

$$\left[\frac{\partial\phi}{\partial y}\right]_{y=0}=\frac{A\cos\pi x'/2s}{(\sin^2\pi c/2s-\sin^2\pi x'/2s)^{\frac{1}{2}}}, \quad 0<x'<c,$$

or, in notation corresponding to Fig. 1,

$$\left[\frac{\partial\phi}{\partial r}\right]_{r=a}=\frac{A\cos\pi z/2s}{(\sin^2\pi c/2s-\sin^2\pi z/2s)^{\frac{1}{2}}}, \quad 0<z<c,$$

that is with c, and hence z, small compared to s,

$$\left[\frac{\partial\phi}{\partial r}\right]_{r=a}\simeq A'[1-(z/c)^2]^{-\frac{1}{2}}, \quad c\ll s, \quad 0<z<c, \tag{4}$$

where A' is a constant $(A'=2sA/\pi c)$.

In order to evaluate the constant A' there is available the expression for the flux Q_1 entering the tube between $z=0$ and $z=c$. This flux is

$$Q_1=2\pi aK\int_0^c \partial\phi/\partial r\Big|_{r=a}\,dz,$$

where K is the soil permeability[4] (thermal conductivity in the heat problem, ϕ being then the temperature). Substituting Eq. (4) in the last expression and integrating, there results

$$Q_1=ac\pi^2A'K.$$

It is more convenient to consider the flow into unit length of drain tube including both the open and the impervious portions of surface. Let Q be this flux. Then, $Q=Q_1/s=ac\pi^2A'K/s$; that is

$$A'=Qs/ac\pi^2K \tag{5}$$

and Eq. (4) becomes

$$[\partial\phi/\partial r]_{r=a}\simeq(Qs/ac\pi^2K)[1-(z/c)^2]^{-\frac{1}{2}}, \quad 0<z<c. \tag{6}$$

This expression, since we took $b>s$, does not involve b.

As has been indicated, the right-hand sides of Eqs. (2) and (6) require development in a Fourier series. A cosine series is needed. It is

$$[\partial\phi/\partial r]_{r=a}=\sum_{n=1}^{\infty}A_n\cos n\pi z/s+A_0/2, \quad 0<z<s,$$

where the coefficients are found in the usual way to have the values,

$$A_0/2=Q/2a\pi K, \tag{7}$$

$$A_n=(2Q/a\pi^2K)\sin^{-1}[(s/n\pi c)\sin n\pi c/s]. \tag{8}$$

If we now define $R_0(n\pi r/s)$ by

$$R_0(n\pi r/s)=\frac{K_0(n\pi b/s)I_0(n\pi r/s)-I_0(n\pi b/s)K_0(n\pi r/s)}{K_0(n\pi b/s)I_1(n\pi a/s)+I_0(n\pi b/s)K_1(n\pi a/s)}, \tag{9}$$

in which I_0 and K_0 are modified Bessel functions of the

[2] Two-dimensional boundary conditions in solving problems of axial symmetry have recently been applied by W. R. Smythe, Rev. Mod. Phys. 20, 176 (1948).

[3] W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939), first edition, p. 89.

[4] Muskat (see reference 1, p. 294) uses the symbol k where we use K. K is the quantity of water seeping through a unit length of soil of unit cross section when the difference in hydraulic head across the unit length is unity. Our ϕ has the physical significance of hydraulic head and is equivalent to Muskat's Φ/k.

first, and second kind and zero order; and I_1, of the first kind and first order; we can write down a function ϕ which satisfies $\nabla^2\phi=0$ and conditions (II), (III), (IV), and (V) exactly, and (I) approximately. It is

$$\phi=\sum_{n=1}^{\infty}\frac{A_n s}{n\pi}R_0\left(\frac{n\pi r}{s}\right)\cos\frac{n\pi z}{s}+\frac{A_0}{2}a\ln\frac{r}{b}+V_b, \quad (10)$$

where the A_n and A_0 are given by Eqs. (7) and (8).

The right-hand side of Eq. (10) does not converge rapidly; nor has an analytic summation for it been found. Furthermore the flux values based on Eq. (10), as numerical calculations show, are not appreciably different for practical purposes, from the values resulting from a simpler expression obtained as follows.

SIMPLIFIED APPROXIMATE SOLUTION OF SPECIFIED UNCERTAINTY

Let (instead of Eq. (4))

$$[\partial\phi/\partial r]_{r=a}=B', \quad 0<z<c; \quad (11)$$

then, as Eq. (5) was determined, one finds

$$B'=Qs/2ac\pi K.$$

Also, developing the right-hand side of Eqs. (1) and (11) into a Fourier cosine series in the interval $0<z<s$, there results

$$[\partial\phi/\partial r]_{r=a}=\sum_{n=1}^{\infty}B_n\cos\frac{n\pi z}{s}+\frac{B_0}{2},$$

where

$$B_0/2=cB'/s,$$
$$B_n=(2B'/n\pi)\sin n\pi c/s.$$

Instead of Eq. (10) we now have

$$\phi=\sum_{n=1}^{\infty}(B_n s/n\pi)R_0(n\pi r/s)\cos n\pi z/s+(B_0/2)a\ln r/b+V_b;$$

or, in terms of the flux Q,

$$\phi=\frac{Q}{2\pi K}\left[\ln\frac{r}{b}-\frac{2s^2}{ac\pi^2}S(r,z)\right]+V_b \quad (12)$$

where

$$S(r,z)=-\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\frac{n\pi c}{s}R_0\left(\frac{n\pi r}{s}\right)\cos\frac{n\pi z}{s}. \quad (13)$$

In Eq. (12) $\phi(r,z)$ is given in terms of supposedly known values of V_b and Q. If, however, $\phi(r,z)$ is known for some point (r,z), the flux Q may be considered as unknown and there results

$$Q=\frac{2\pi K[V_b-\phi(r,z)]}{\ln b/r+(2s^2/ac\pi^2)S(r,z)}. \quad (14)$$

We will now establish, referring to Fig. 2, lower and upper limits Q_l and Q_u for the flux. In the figure the line EF represents the equipotential passing through the point $(r,z)=(a,c)$, the point E being at higher potential than the point A. Now, since in the actual physical problem the potential along AE does not increase but has the constant value V_a, identification of V_a with $\phi(r,z)=\phi(a,0)$ in Eq. (12) establishes a lower limit Q_l of flow.

Suppose now that V_a is identified with the equipotential EF,[5] specified analytically by $\phi=\phi(a,c)$. Then, in order that the potential along AE be constant, a section AEF of soil must be removed. But with this soil removed the flow will be greater than if the potential V_a were applied along AE with no soil removed. Therefore, identification of $\phi(r,z)$ with $\phi(a,c)=V_a$ yields an upper limit of flow.

From the above considerations and Eqs. (13) and (14) we have then

$$Q_l=\frac{2\pi K(V_b-V_a)}{\ln b/a+(2s^2/ac\pi^2)S(a,0)}, \quad (15)$$

$$Q_u=\frac{2\pi K(V_b-V_a)}{\ln b/a+(2s^2/ac\pi^2)S(a,c)}; \quad (16)$$

and we define an average value of flux Q_{Av} by

$$Q_{Av}=(Q_l+Q_u)/2, \quad (17)$$

a value which should not be far in error.

Although Eqs. (15) and (16) are correct lower and upper limits of the flow for all values of a, b, c, and s, the expressions may be simplified. This is true, because for

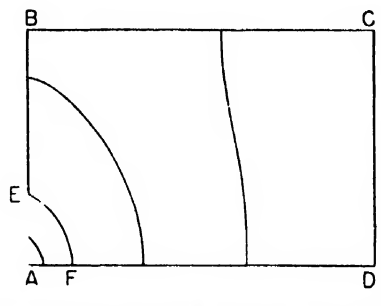


FIG. 2. Schematic equipotentials for the region $ABCD$ of Fig. 1 when the normal component of flux density across the stricture AE is constant.

the range of values of a, b, and s considered in the following, there is valid the approximation $R_0(n\pi a/s)=-K_0(n\pi a/s)/K_1(n\pi a/s)$ to better than 1 part in 10^8. (The approximation $R_0(n\pi r/s)=-K_0(n\pi r/s)/K_1(n\pi a/s)$ is correct to better than 1 part in 10^5, the worst case being for $n=1$ and $r=b$.) There results then, finally, for

[5] The distance AF in Fig. 2 has been computed, using Eq. (10). For $2a=19.2/\pi\simeq 6$, $b=48$, $2c=4/15\simeq\frac{1}{4}$, $s=6$, it is $AF=0.075$, where, for a practical example all units may be taken as inches. It will be shown that $b=48$ in. corresponds to a depth of 24 in. for a horizontally installed drain tube.

numerical calculations

$$Q_l=\frac{2\pi K(V_b-V_a)}{\ln b/a+(2s^2/ac\pi^2)S_1}, \tag{18}$$

$$Q_u=\frac{2\pi K(V_b-V_a)}{\ln b/a+(2s^2/ac\pi^2)S_2}, \tag{19}$$

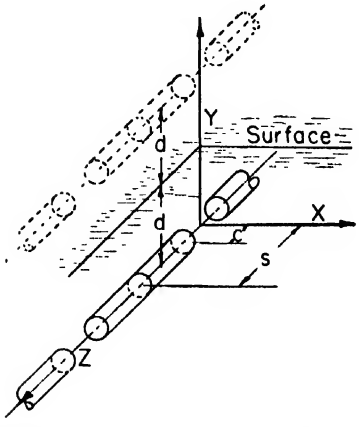


FIG. 3. Image case for a single drain tube. The semi-distance c between impermeable units is indicated larger than is found in practice.

where

$$S_1=S(a,0)=\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\frac{n\pi c}{s}\frac{K_0(n\pi a/s)}{K_1(n\pi a/s)}, \tag{20}$$

$$S_2=S(a,c)=\sum_{n=1}^{\infty}\frac{1}{2n^2}\sin\frac{2n\pi c}{s}\frac{K_0(n\pi a/s)}{K_1(n\pi a/s)}. \tag{21}$$

A commonly recommended size of drain tube for laterals (as distinguished from mains) is 5 in. inside diameter. Smaller diameters, sometimes as small as 2 in. have been used. Unit lengths are almost invariably 1 ft., and wall thickness about $\frac{1}{2}$ in. In view of the above and in order to simplify computations the following values for geometrical quantities have been used: $a/s=0.8/\pi$ and $1.6/\pi$; $b/s=8$, 16 and 200; $c/s=1/1440$, $2/1440$, $\cdots$, $32/1440$. With $2s=1$ ft. the values of a/s and c/s correspond to $2a=3.06$ and 6.12 in., and $2c$ to approximately $1/128$, $1/64$, $\cdots$, $\frac{1}{4}$ in. The values chosen for b/s correspond (with $2s=1$ ft.) to $b=4$, 8 and 100 ft.

Numerical results for S_1 and S_2 are presented in Table I. The values were computed from Eqs. (20) and (21) by summing term by term, from $n=1$ to $n=44$ and using Euler-McLaurin[6] summation formulas for $n=45$ to ∞, taking for $n\geq 45$, $K_0(n\pi a/s)/K_1(n\pi a/s)=1$. Taking the latter ratio = 1 accounts for the uncertainty in the last place of decimals in Table I. In preparing Table I the relation $S_2[(1/2)c/s]=(1/2)S_1(c/s)$ was useful. There is a similar relation for Q_l and Q_u.

[6] Cf. E. T. Whittaker and G. Robinson, *The Calculus of Observations* (Blackie and Son, Limited, London, 1924), p. 135.

TABLE I. Values of S_1 and S_2, Eqs. (20) and (21).

	$a/s=1.6/\pi$		$a/s=0.8/\pi$	
c/s	S_1	S_2	S_1	S_2
32/1440	0.22912 (2)[a]	0.18186 (1)	0.21141 (4)	0.16515 (1)
16/1440	0.13849 (2)	0.11455 (1)	0.12942 (4)	0.10572 (2)
8/1440	0.08128 (4)	0.06925 (1)	0.07668 (9)	0.06471 (2)
4/1440	0.04666 (4)	0.04066 (2)	0.04434 (9)	0.03834 (5)
2/1440	0.02634 (4)	0.02333 (2)	0.02517 (6)	0.02217 (4)
1/1440	0.01468 (2)	0.01317 (2)	0.01405 (4)	0.01259 (3)

[a] Numbers in parentheses indicate uncertainty (±) in the fifth decimal place.

Table II presents values of Q_l', Q_M' and Q_u', where,

$$\left.\begin{aligned} Q_l'&=Q_l/K(V_b-V_a),\\ Q_M'&=Q_M/K(V_b-V_a),\\ Q_u'&=Q_u/K(V_b-V_a).\end{aligned}\right\} \tag{22}$$

The uncertainty in Q_M' is greatest, $51/570=8.95$ percent for the values $b/s=8$, $a/s=0.8/\pi$ and $c/s=32/1440$; and (except for $c/s=1$ when the uncertainty is zero) is least, $14/437=3.20$ percent for the values $b/s=200$, $a/s=1.6/\pi$, and $c/s=1/1440$. The uncertainty in Q_M' increases with decreasing a/s and b/s and increasing c/s. The uncertainty 3.20 percent found for $c/s=1/1440$ is larger than was expected.

IMAGE CASE: A SINGLE DRAIN TUBE

Figure 3 represents the image case. The drain tube is buried horizontally (nearly, in practice) at depth d below a plane water table, here taken as the surface of

TABLE II. Values of Q_l', Q_M', and Q_u'.[a]

	$a/s=1.6/\pi$			$a/s=0.8/\pi$		
c/s	Q_l'	Q_M'	Q_u'	Q_l'	Q_M'	Q_u'
			$b/s=8$			
1	2.281	2.281	2.281	1.823	1.823	1.823
32/1440	0.916	0.981	1.045	0.570	0.621	0.671
16/1440	0.815	0.865	0.916	0.494	0.532	0.570
8/1440	0.733	0.774	0.814	0.435	0.465	0.494
4/1440	0.666	0.699	0.732	0.389	0.412	0.435
2/1440	0.610	0.638	0.666	0.351	0.370	0.389
1/1440	0.563	0.586	0.610	0.321	0.336	0.352
			$b/s=16$			
1	1.823	1.823	1.823	1.518	1.518	1.518
32/1440	0.832	0.885	0.937	0.537	0.581	0.625
16/1440	0.747	0.790	0.832	0.469	0.503	0.536
8/1440	0.678	0.713	0.747	0.415	0.442	0.469
4/1440	0.620	0.649	0.678	0.373	0.394	0.415
2/1440	0.572	0.596	0.620	0.338	0.356	0.373
1/1440	0.530	0.551	0.572	0.310	0.324	0.338
			$b/s=200$			
1	1.052	1.052	1.052	0.943	0.943	0.943
32/1440	0.624	0.652	0.681	0.441	0.470	0.499
16/1440	0.573	0.599	0.624	0.394	0.418	0.441
8/1440	0.533	0.554	0.575	0.356	0.375	0.394
4/1440	0.496	0.515	0.533	0.324	0.340	0.356
2/1440	0.465	0.481	0.496	0.298	0.311	0.324
1/1440	0.437	0.451	0.465	0.276	0.287	0.298

[a] Q_l', Q_M' and Q_u' are respectively a lower, an average, and an upper limit of flow for a unit length of a long drain tube of the type indicated in Fig. 1, the permeability of the conducting medium, and the difference in potential across the outer and inner flow boundaries (IV and I of Fig. 1), being unity. Cf. Eqs. (18), (19), and (22).

TABLE III. Relative flow into cracks between 1 ft. long[a] impervious drain tube units.

Crack width inches	Depth of tube 2 ft. Diameter 6 in.	Depth of tube 2 ft. Diameter 3 in.	Depth of tube 4 ft. Diameter 6 in.	Depth of tube 4 ft. Diameter 3 in.
Flow into a completely open[b] drain tube taken as unity				
1/128	0.257	0.185	0.302	0.214
1/64	0.280	0.203	0.327	0.234
1/32	0.306	0.226	0.356	0.260
1/16	0.339	0.255	0.391	0.291
1/8	0.379	0.292	0.433	0.331
1/4	0.430	0.341	0.485	0.383
Inf.	1.000	1.000	1.000	1.000
Flow into a crack 1/32 in. wide taken as unity				
1/32	1.000	1.000	1.000	1.000
1/16	1.100	1.128	1.099	1.116
1/8	1.238	1.290	1.216	1.272
1/4	1.405	1.508	1.361	1.432
Inf.	3.268	4.425	2.809	3.846

[a] Crack widths, tube diameters and length of tube units are nominal. The entries in the table correspond actually (see Fig. 3) to $s=6$ in. (not $s-c=6$ in.) to $2c=1/120$, 1/60, 1/30, etc. in. and to $2a=19.2/\pi=6.06$ in. and $2a=9.6/\pi=3.03$ in. Infinite crack width corresponds to $c/s=1$. The depths 2 and 4 ft. are correct.

[b] For practical purposes in soil drainage, a drain tube embedded in gravel

the soil. The soil is of uniform permeability K and extends downward to great depth, the drain tube being the only exit for soil water. The fictitious image drain is distance d above the soil surface. With the coordinate axes as shown, and referring to Eq. (12), the potential θ may be written down by inspection and is

$$\theta=\frac{F}{2\pi K}\left\{\ln\frac{(x^2+y^2)^{\frac{1}{2}}}{[x^2+(2d-y)^2]^{\frac{1}{2}}}+\frac{2s^2}{ac\pi^2}\sum_{n=1}^{\infty}G_n n^{-2}\sin n\pi c/s\cdot\cos n\pi z/s\right\}+V_d, \quad (23)$$

where

$$G_n=R_0\left[\frac{n\pi}{s}(x^2+y^2)^{\frac{1}{2}}\right]-R_0\left\{\frac{n\pi}{s}[x^2+(2d-y)^2]^{\frac{1}{2}}\right\},$$

and where F, now used instead of Q, is the flux per unit length of drain tube, including the open and closed parts, and V_d is the hydraulic head at the surface of the soil.

We assume that the drain line is running full with zero back pressure or suction. Therefore, taking the axis of the drain tube as reference level for hydraulic head, the potential V_w over the inside of the tube is $V_w=a$; also $V_d=d$ if the soil is water-saturated to the surface; and $V_d=d+t$, if in addition there is a surface layer of water of thickness t.

Let us consider Eq. (23) for the case that $c=s$, that is, for the case that the summation is zero. Then, taking $x=0$ and $y=a$, there results $F=2\pi K(V_d-V_w)/\ln(2d-a)/a$; and taking $x=0$ and $y=-a$, there results $F=2\pi K(V_d-V_w)/\ln(2d+a)/a$. One of these values is too high and one too low.[7] A value which more nearly equals the actual flux is $F=2\pi K(V_d-V_w)/\ln 2d/a$.[8] Similarly, a best value for R_0-terms at the surface of the drain tube is

$$R_0(n\pi a/s)-R_0(2n\pi d/s)=-[K_0(n\pi a/s)/K_1(n\pi a/s)-K_0(2n\pi d/s)/K_1(n\pi a/s)],$$

very nearly, $=-[K_0(n\pi a/s)/K_1(n\pi a/s)]$, very nearly (for practical values of d, s and a). Therefore Eq. (23) may be written for the drain surface, upon solving for F,

$$F=2\pi K(V_d-V_w)\{\ln 2d/a+(2s^2/ac\pi^2)\sum_{n=1}^{\infty}n^{-2}\sin n\pi c/s\cdot\cos n\pi z/s\cdot K_0(n\pi a/s)/K_1(n\pi a/s)\}^{-1}.$$

Taking now, as was done in obtaining Eqs. (18) and (19), $z=0$ and then $z=c$, lower and upper limits, F_l and F_u, for the flux are found to be

$$F_l=\frac{2\pi K(V_d-V_w)}{\ln 2d/a+(2s^2/ac\pi^2)S_1}, \quad (24)$$

$$F_u=\frac{2\pi K(V_d-V_w)}{\ln 2d/a+(2s^2/ac\pi^2)S_2}, \quad (25)$$

where S_1 and S_2 are given by Eqs. (20) and (21). Also, as in Eq. (17), we define

$$F_{av}=(F_l+F_u)/2; \quad (26)$$

and as in Eq. (22),

$$\left.\begin{aligned}F_l'&=F_l/K(V_d-V_w)\\F_{av}'&=Q_{av}/K(V_d-V_w)\\F_u'&=F_u/K(V_d-V_w)\end{aligned}\right\}. \quad (27)$$

Comparing Eqs. (18) and (19) with Eqs. (24) and (25), and Eq. (17) with Eq. (26), it is now apparent that Table II gives values of F_l', F_{Av}' and F_u', if b is taken

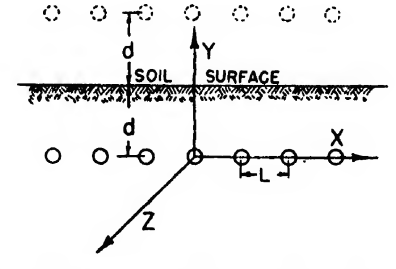


FIG. 4. A series of equally spaced drain tubes.

[7] The former value is too high, because the equipotential surface $\theta(0, a)$ lies outside the actual drain tube except at $(0, a)$; the latter, too small, because the equipotential $\theta(0, -a)$ lies inside the actual drain tube except at $(0, -a)$.

[8] The exact value (see reference 3, p. 77, Eq. 1) is $F=2\pi K(V_d-V_w)/\cosh^{-1}d/a$. The expression $\ln 2d/a$ approximates $\cosh^{-1}d/a$ to better than 1 part in 3400 for $d=4$ and $a=0.25$.

equal to $2d$, and if F's are substituted for Q's. With these considerations Table III has been prepared. The table brings out the fact that the width of opening between drain pipe units is an important factor in seepage calculations. For example, in a clay soil, where a crack width of $\frac{1}{4}$ in. may safely be used, flow to a 6 in. tube at 4 feet depth may be increased 36 percent over that achieved by putting the drain pipe units $\frac{1}{32}$ in. apart as is commonly done; while embedding the tubes in gravel may increase the flow 180 percent. In soils where gravel is not available and where wide drain tube openings might result in the entry of soil into the tubes, with possible clogging, screens could be placed over the openings. The soil would be stabilized by the time the screens deteriorate.

IMAGE CASE: MULTIPLE DRAIN TUBES

For a series of drain tubes equally spaced at distance L and at depth d (Fig. 4), the potential is, comparing Eq. (23),

$$\theta=\frac{F}{2\pi K}\sum_{m=-\infty}^{\infty}\left[\frac{1}{2}\ln\frac{(x+mL)^2+y^2}{(x+mL)^2+(2d-y)^2}+T_m\right]+V_d, \quad (28)$$

where

$$T_m=\frac{2s^2}{ac\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\frac{n\pi c}{s}\cos\frac{n\pi z}{s}\times\left(R_0\{(n\pi/s)[(x+mL)^2+y^2]^{\frac{1}{2}}\}-R_0\{(n\pi/s)[(x+mL)^2+(2d-y)^2]^{\frac{1}{2}}\}\right).$$

Now the first summation in Eq. (28) is[9]

$$(1/2)\ln(\cosh 2\pi y/L-\cos 2\pi x/L)/(\cosh 2\pi(2d-y)/L-\cos 2\pi x/L);$$

or putting $y=a$, $x=0$, it is

$$(1/2)\ln(\cosh 2\pi a/L-1)/[\cosh 2\pi(2d-a)/L-1]=\ln[(\sinh\pi a/L)/\sinh\pi(2d-a)/L].$$

The second summation in Eq. (28), when $y=a$ and $x=0$, reduces, since $d\gg s$, $L\gg s$, to the single summation $\sum_{n=1}^{\infty}[-K_0(n\pi a/s)/K_1(n\pi a/s)]n^{-2}\sin n\pi c/s\cdot\cos n\pi z/s$, in which $(-K_0/K_1)$ has been written, as before, for R_0. Therefore, Eq. (28) yields, taking first $x=0$, $y=a$, $z=0$, and then $x=0$, $y=a$, $z=c$, lower and upper limits for the flux as follows.

$$F_l=\frac{2\pi K(V_d-V_w)}{\ln[(\sinh 2\pi d/L)/(\sinh\pi a/L)]+(2s^2/ac\pi^2)S_1}, \quad (29)$$

$$F_u=\frac{2\pi K(V_d-V_w)}{\ln[(\sinh 2\pi d/L)/(\sinh\pi a/L)]+(2s^2/ac\pi^2)S_2}, \quad (30)$$

where the a in $\sinh\pi(2d-a)/L$ has been dropped for the same reason that the a in $(2d-a)$ was dropped in obtaining Eqs. (24) and (25).

It is apparent, comparing Eqs. (18), (19); (24), (25); and (29), (30), that the same two correction terms $(2s^2/ac\pi^2)S_1$ and $(2s^2/ac\pi^2)S_2$ will apply in a number of drainage problems.[10,11] These terms, multiplied by $1/2\pi K$, represent upper and lower limits of a resistance term which, in order to compute the flux, must be added to the resistance of the medium when calculated for completely open drain tubes. Stated in another way, the loss in hydraulic head due to strictures in a drain tube of the type shown in Fig. 1, whether there is a single tube or a number of tubes in an installation, will not exceed $(Q/2\pi K)(2s^2/ac\pi^2)S_1$; nor be less than $(Q/2\pi K)(2s^2/ac\pi^2)S_2$, where Q is the flow per unit length of tube.

[9] Compare Muskat, reference 1, p. 190.

[10] Don Kirkham, "Flow of ponded water into drain tubes in soil overlying an impervious layer," Trans. Amer. Geophys. Union **30**, 369–385, 1949.

[11] Y. Gustaffon, Untersuchungen Über die stromungsverhältnisse in gedräntem Boden, Acta. Agr. Suecana, Stockholm **2**, No. 1 (1946).

Determination of Modified Index-of-Refraction over the Gulf of Mexico from Radio Data*

A. W. Straiton and A. H. LaGrone
The University of Texas, Austin, Texas
(Received November 21, 1949)

This paper presents the results of measurements of the signal strength and phase of 3.2-centimeter radio waves for a 26.5-mile path over the Gulf of Mexico under a variety of meteorological conditions. The radio data are used to determine the modified index-of-refraction as a function of height and the associated attenuation factor. The modified index-of-refraction curves are compared to those measured meteorologically.

I. INTRODUCTION

THE strength of radio waves received through the troposphere is influenced by the distribution of the index-of-refraction with height. For this work a modified index-of-refraction, M,** is used to permit the use of a plane earth representation. Numerous papers have described the theory and measurements from which the radio signal strength may be determined from a known distribution of index with height.

Macfarlane[1] pointed out the possibility of reversing the process so as to obtain the index distribution with height from radio data. His method requires a knowledge of the height-gain curve and additional data from which the wave attenuation with distance may be determined.

Macfarlane's method may be extended[2] to use as known data the phase distribution with height as well as the signal strength distribution. This eliminates the necessity of additional information from which the attenuation may be determined. This modified method is used in this paper to obtain the low level index-of-refraction distribution on a 26.5-mile over-water path using a wave-length of 3.2 centimeters. These data, taken during June and July, 1949, are associated with a variety of meteorological conditions ranging from an almost standard atmosphere to very strong ducts.

II. RADIO MEASUREMENTS

The radio path was over a section of the Gulf of Mexico adjacent to Bolivar Peninsula near Galveston, Texas as shown in Fig. 1. A 3.2-centimeter transmitter was located on a 50-foot tower on the mainland beach near High Island, Texas. A 2K39 Klystron supplied a continuous wave output of approximately 200 milliwatts. A picture of the transmitting tower is shown in Fig. 2.

Radio receivers were located on the northeast point of Galveston Island on two towers. The phase-difference equipment developed by the Electrical Engineering Research Laboratory of The University of Texas[3] was located on the elevator of a 60-foot tower with its base at 10 feet above mean sea level as shown in Fig. 3. This permitted measurements of the signal strength and phase difference from 15 to approximately 60 feet mean sea level. Three antennas were mounted on the elevator in a vertical line with a spacing of five feet between adjacent antennas. The signal strength was measured at the top and bottom antennas, and the phase difference was measured between the center and the two outside antennas. This permitted duplicate signal-strength and phase data over most of the height range obtained with the 60-foot tower. An auxiliary 16-foot tower with a movable elevator, as shown in Fig. 4, permitted measurement of signal strength from 2 to approximately 15 feet above the mean sea level. Horizontal polarization was used for the test.

III. METEOROLOGICAL MEASUREMENTS

Meteorological stations were set up at two points along the radio path. One station was located at a lighthouse, Station B in Fig. 1. Wet- and dry-bulb temperature soundings were made over a range of heights up to 60 feet, and wind speed and direction measurements were made at a height of about 70 feet.

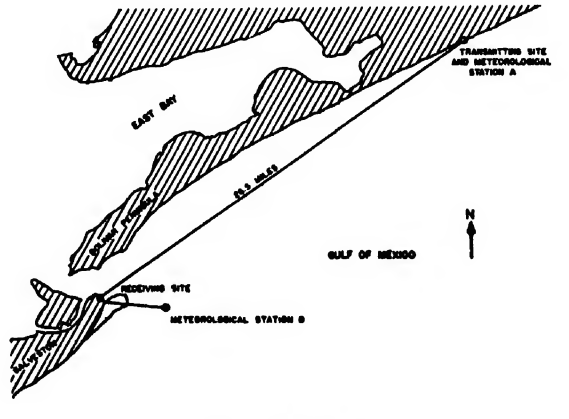


FIG. 1. Radio path.

* This work was sponsored by the ONR at The University of Texas under Contract N5ori-136, P.O.I.

** $M=(79/T)(P+4800e/T)+0.048h$, where T, P, e, and h are temperature in degrees Kelvin, atmospheric pressure in millibars, vapor pressure in millibars, and height in feet. This permits the use of a plane earth representation. For the standard atmosphere, $dm/dh=0.036$ M units per foot.

[1] G. G. Macfarlane, "A Method of Deducing the Refractive-Index Profile of a Stratified Atmosphere from Radio Observations," Meteorological Factors in Radio Wave Propagation; report of a conference of The Physical Society, London, 1948.

[2] A. W. Straiton, J. App. Phys. 20, 228 (1949).

[3] F. E. Brooks, Jr. and C. W. Tolbert, "A Receiver for Measuring Angle-of-Arrival in a Complex Wave," Report No. 33, Electrical Engineering Research Laboratory, The University of Texas, 1949.

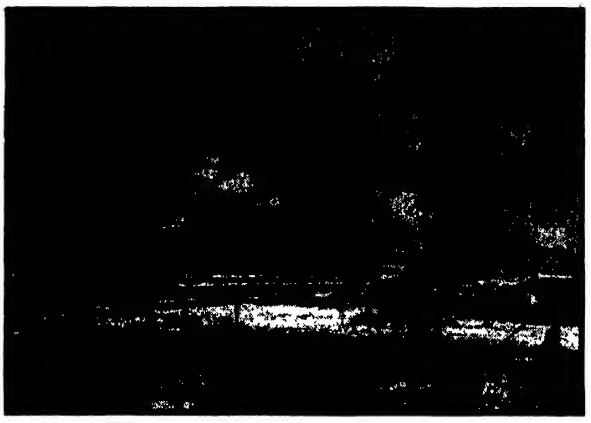


FIG. 2. Transmitting tower looking toward receiver.

The second meteorological station was on an oil drilling platform, Station A, approximately 50 feet off shore near the transmitter. Wet- and dry-bulb temperature measurements were made up to heights of 60 feet, and wind speed and direction were measured at the 10-foot platform level. A detailed description of the meteorological stations and the measurements is given elsewhere.[4] The meteorologically-measured modified index of refraction taken simultaneously with the radio data will be shown in this report for comparison with those determined from radio data.

IV. THEORETICAL CONSIDERATIONS

For radio propagation well beyond the line of sight, the electrical field vector, ψ, may be expressed as the sum of a number of normal modes[5] as follows:

$$\psi = e^{i\omega t - i\pi/4}(2\pi/kr)^{\frac{1}{2}} \sum_m e^{-\gamma_m r} U_m(h_1) U_m(h), \qquad (1)$$

where r is the distance around the earth's surface, k is 2π divided by the wave-length, γ_m is the propagation function for the mth mode, $U_m(h_1)$ and $U_m(h)$ are the height-gain function for the transmitter and receiver, respectively, for the mth mode.

For the distances and heights used in the measurement program, all modes higher than the first may usually be neglected.[6] The height-gain function is then given for the first mode by:

$$(d^2U_1/dh^2) + k^2[y(h) + \Lambda_1]U_1 = 0, \qquad (2)$$

where Λ_1 is characteristic value for the first mode.

V. MACFARLANE'S EQUATION

Macfarlane[1] expressed Eq. (2) as follows: Let

$$U_1 = Ae^{i\phi} \quad \text{and} \quad \Lambda_1 = a + ib.$$

Then

$$k^2(y+a) = (\dot{\phi})^2 - d^2A/dh^2/A \qquad (3)$$

and

$$\dot{\phi} = \frac{-k^2 b \int_0^h A^2 dh}{A^2} \qquad (4)$$

where the dots indicate derivatives with respect to h.

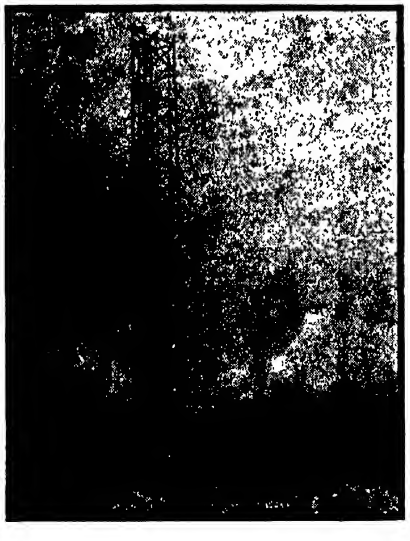


FIG. 3. Receiving tower with phase-difference equipment.

FIG. 4. Auxiliary tower for measurement of low level signal strengths.

[4] J. R. Gerhardt, "Summary of Meteorological Measurements Made for an Overwater Path near Galveston, Texas, during June and July, 1949," Report No. 36, Electrical Engineering Research Laboratory, The University of Texas, 1949.

[5] W. H. Furry, "Theory of Characteristic Functions in Problems of Anomalous Propagation," Report No. 680, Radiation Laboratory, Massachusetts Institute of Technology, 1945.

[6] H. W. Smith and A. W. Straiton, "Three Centimeter Radio Wave Propagation in a Surface Duct over The Gulf of Mexico," Report No. 26, Electrical Engineering Research Laboratory, The University of Texas, 1949.

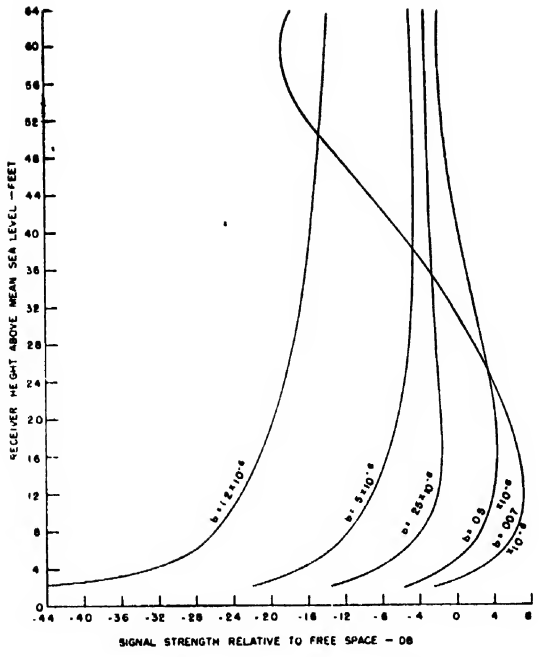


FIG. 5. Measured signal-strength curves on a db basis relative to free space.

Macfarlane replaced ϕ in Eq. (3) by the right side of Eq. (4). Then, if the attenuation factor b is known from measurements at other frequencies or distance, $(y+a)$ can be determined directly.

However, since the measurements made at the Electrical Engineering Research Laboratory were of A and $\dot{\phi}$, $(y+a)$ can be determined directly from (3), and b from (4).

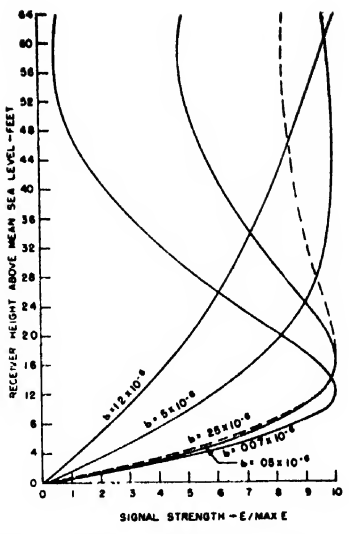


FIG. 6. Measured signal-strength curves relative to maximum signal in height interval.

VI. PROCEDURE FOR FINDING M CURVES

The procedure used was as follows:

1. The phase difference between the signals at the receivers and the signal strength at the receivers were obtained from the measurement charts. This provided the signal strength, A, and the phase difference, $\dot{\phi}$, both as a function of height.
2. The first derivative of A with respect to h was obtained graphically and plotted as a function of h. A smoothed first derivative was drawn and the original A curve was modified in accordance with the smoothed first derivative curve.
3. The second derivative with respect to h, was found graphically and divided by A.
4. A smooth curve was drawn through the values of ϕ as a function of height.
5. The values of $(y+a)=[2(M-M_0)\times 10^{-6}+a]$ were found from Eq. (3).
6. Curves of $\dot{\phi}$ for several assumed values of b were plotted

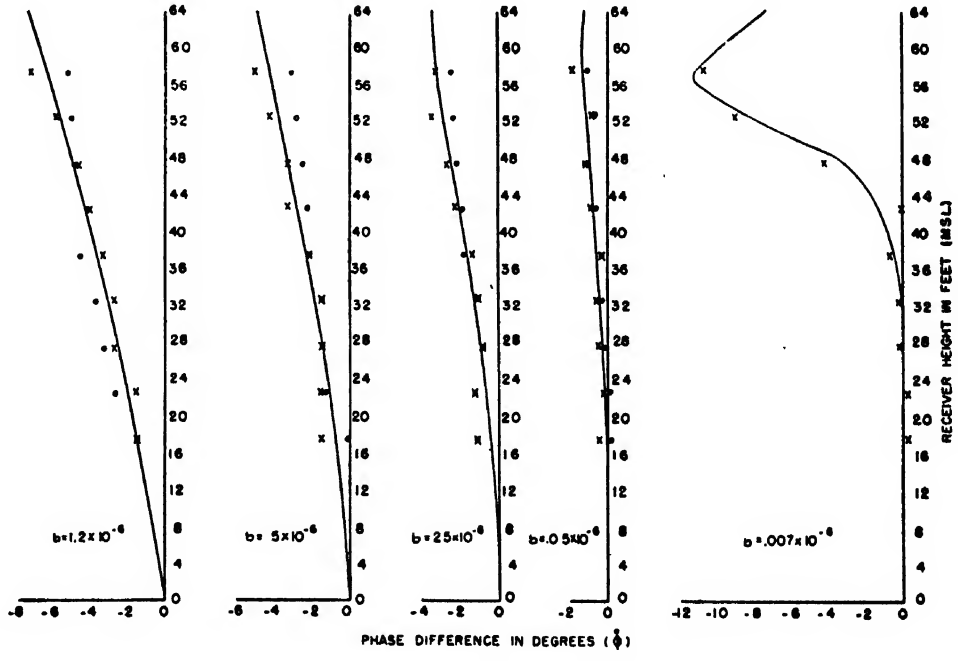


FIG. 7. Measured phase-difference curves.

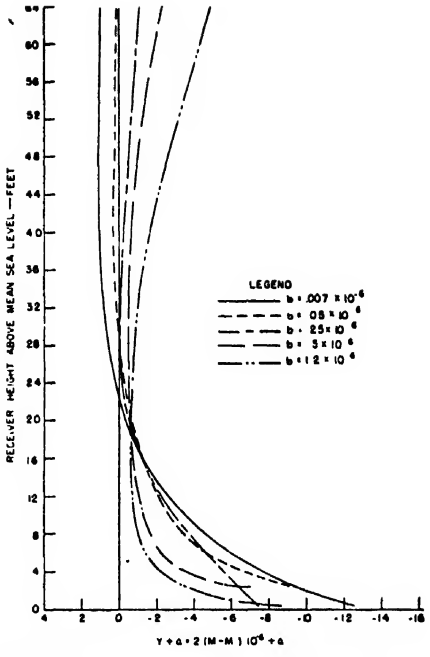


FIG. 8. M curves determined from radio data.

from Eq. (4). When the measured values of ϕ were superimposed on these curves, the value of b was chosen such that the measured curve coincided as nearly as possible with the one determined from the A data.

VII. HEIGHT-GAIN CURVES

In measuring the height-gain curves, data were taken in several ways. The receiver on the auxiliary 16-foot tower was moved up and down while its received signal was being recorded. The 16-foot tower receiver was then set at 15 feet and left there as a reference while the main receiver on the 60-foot tower was moved up continuously and down in five-foot steps. Since the signal was scintillating rapidly, averages of the various readings were taken to establish the height-gain curves shown. Fifty-two sets of such readings were taken and the height-gain or A curves were drawn. Many of the A curves were near duplicates and others were very similar. The entire group can be well represented by five samples which include the two extreme signals and three intermediate values. The signal strength-height curves for these five cases are shown on a db basis relative to free space in Fig. 5. The same curves are shown in Fig. 6 in terms of the field strength at a given height divided by the maximum field strength in the interval included. The curves representing the various examples are the ones which gave the smoothed first derivative curves. Seven intermediate values are added to increase to twelve the number used for the tabulation shown later.

TABLE I. Phase-integral condition.

Number	Date	Time	$\int_0^{Z_0^1}(y+a+ib)dh$	$\frac{1}{2}(\pi/k)$
1*	7- 7-49	0530–0647	$0.0346/\underline{-3°}$	$0.0394/\underline{0}$
2*	7- 7-49	0815–0900	$0.0362/\underline{-0.7°}$	$0.0394/\underline{0}$
3*	6-30-49	0043–0105	0.0351	0.0394
4	6-29-49	0550–0621	-0.0379	0.0394
5	7- 6-49	1130–1210	0.0373	0.0394
6	7- 6-49	1235–1300	0.0405	0.0394
7*	6-29-49	1020–1059	0.0399	0.0394
8	6-27-49	1729–1815	0.0410	0.0394
9	6-28-49	0942–0957	0.0410	0.0394
10	6-27-49	1535–1555	0.0412	0.0394
11	6-28-49	0830–0900	0.0402	0.0394
12*	6-27-49	1920–2001	0.0395	0.0394

* Cases shown in graphs.

VIII. PHASE DIFFERENCE-HEIGHT CURVES

The phase-difference curves corresponding to the signal-strength curves shown in Figs. 5 and 6 are shown in Fig. 7. The points indicated are the values measured between the center and each of the two outside antennas for the five-foot interval runs. The lines shown are the values of ϕ determined from the corresponding signal-strength curves by Eq. (4). Because of the physical location of the receivers, it was not convenient to use a reference angle. The reference angles were then chosen so that ϕ would fulfill the boundary condition of being zero at zero height.

IX. M CURVES FROM RADIO DATA

The A curves of Fig. 6 and the ϕ curves of Fig. 7 were used to determine $(y+a)$ from Eq. (3). The M curves thus obtained for the five cases considered are shown in Fig. 8. The inability to separate a from $2(M-M_0)$ is really no disadvantage. The reference value of index of refraction, M_0, is not of much significance and could, without harm, be increased or decreased. However, $(y+a)$ or $[2(M-M_0)\times10^{-6}+a]$ is the quantity that appears in the equations for the height-gain function given in reference 5. The degree to which $(y+a)$ goes negative is a measure of the amount of trapping and the points at which $(y+a)$ is zero are the "turning points," Z_0^m, of the W.K.B. approximation solution.

TABLE II. Attenuation factor.

Number	From M curves $b\times10^{-6}$	From radio data $b\times10^{-6}$	From radio data db/mile
1*	1.37	1.2	1.6
2*	0.59	0.5	0.7
3*	0.17	0.25	0.34
4	0.38	0.2	0.27
5	0.23	0.15	0.21
6	0.15	0.15	0.21
7*	0.06	0.05	0.07
8	0.00	0.03	0.04
9	0.00	0.01	0.01
10	0.00	0.01	0.01
11	0.00	0.008	0.01
12*	0.00	0.007	0.01

* Cases shown in graphs.

TABLE III. Positioning of $(y+a)$ curves.

Number	a From radio data	a From shape of M curve
1*	−1.54	−1.60
2*	−2.70	−2.90
3*	−3.03	−3.14
4	−2.40	−2.30
5	−3.50	−3.42
6	−2.92	−3.00
7*	−2.94	−4.08
8	−5.44	−5.59
9	−5.01	−4.95
10	−6.41	−5.75
11	−5.91	−5.92
12*	−5.71	−6.03

* Cases shown in graphs.

TABLE IV. Free-space values.

Number	db below arbitrary reference: From transmitter and receiver gain	From shape of $(y+a)$ curve	Deviation** db	Attenuation factor $b\times10^{-6}$
1*	33.5	29.5	4.1	1.2
2*	33.5	37.5	− 4.0	0.5
3*	33.5	42.5	− 5.9	0.25
4	33.5	37.7	− 4.2	0.2
5	33.5	44.9	−11.4	0.15
6	33.5	36.6	− 3.1	0.15
7*	33.5	43.1	− 9.6	0.05
8	33.5	41.5	− 8.0	0.03
9	33.5	42.3	− 8.8	0.01
10	33.5	42.7	− 9.2	0.01
11	33.5	43.0	− 9.5	0.008
12*	33.5	43.7	−10.2	0.007

* Cases shown in graphs.
** Minus indicates measured signa less than theoretical.

X. CHECK OF PHASE-INTEGRAL CONDITION

From reference 5, the phase-integral condition is known to be

$$\int_0^{z_{01}} (y+a+ib)^{\frac{1}{2}}dz=\frac{(m-\frac{1}{4})\pi}{k}=\frac{3\pi}{4k}. \qquad (5)$$

For the cases when $(y+a)$ becomes zero, (ib) may be neglected and the integral evaluated numerically from the measured data as the area beneath the portion of the $(y+a)$ curve below the first turning point. For the case when $(y+a)$ did not become zero, the curve was approximated by a linear-exponential expression and a complex turning point established. With this turning point, the integral of Eq. (5) may be evaluated. The value of the phase integral of Eq. (5) for twelve cases is compared with the theoretical value in Table I.

From Table I it is seen that the phase-integral condition is approximately satisfied.

XI. DETERMINATION OF WAVE ATTENUATION

As discussed in Section V, the wave-attenuation factor b may be determined from Eq. (4) by use of radio data. This attenuation is relative to the inverse square root relationship, and may be converted to decibels per mile as shown in Table II.

The attenuation factors may also be determined from the shape of the M curves which were obtained from radio data. The procedure used was to express the M curves from radio data in the form of a linear-exponential; and from the charts given by Pekeris and Ament,[7] values of b were obtained. A comparison of these values of b with those determined by Eq. (4) is also shown in

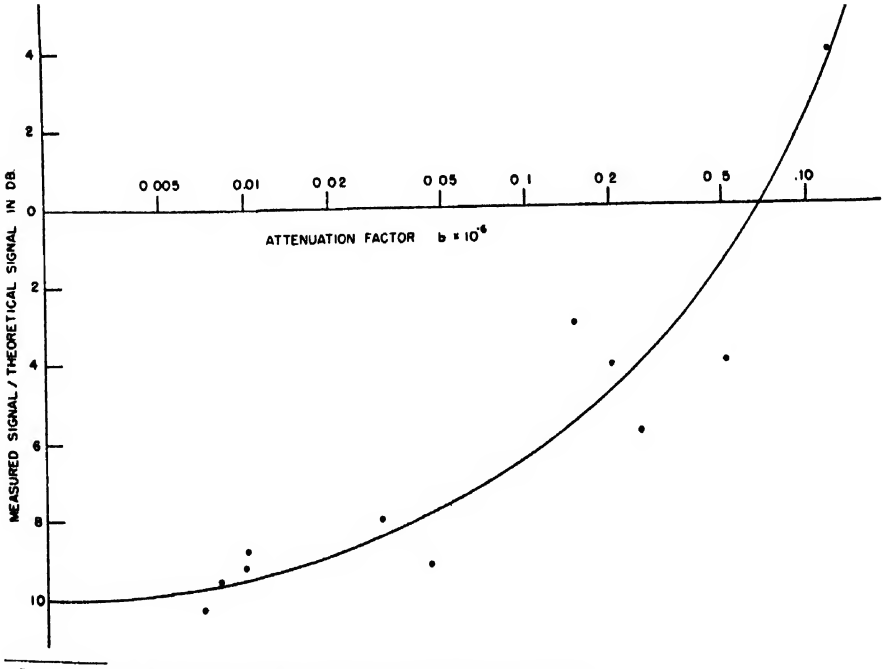


FIG. 9. Comparison of M curves from radio data with meteorological data.

[7] C. L. Pekeris and W. S. Ament, Phil. Mag. 38, 801–823 (1947).

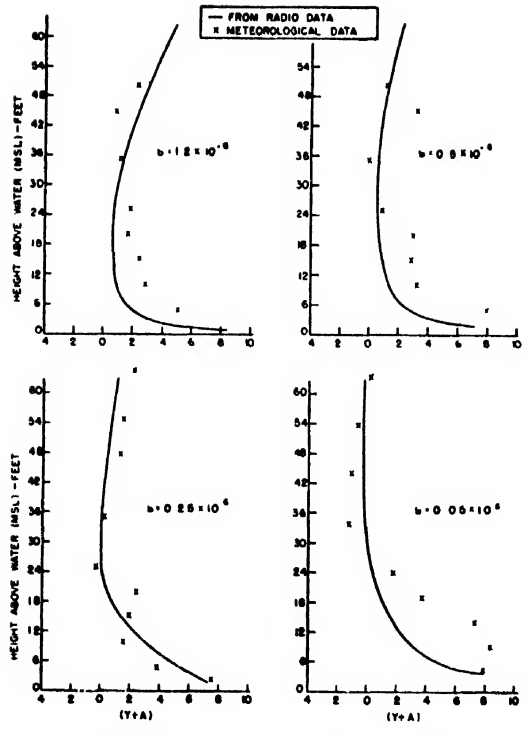


FIG. 10. Relationship between measured and theoretical signal strength.

Table II. The values of b obtained from the charts for cases 8 through 12 are shown as zero, since the scale of the charts was such that the values could not be distinguished from zero. There would, of course, be some attenuation, and the value obtained from radio data is thought to be fairly accurate.

XII. CHECK OF POSITIONING OF M CURVES

The form of linear exponential used by Pekeris and Ament[7] was

$$M-M_0=(0.036h+ge^{-ch}).$$

Then $(y+a)$ will be given by

$$(y+a)=[2(M-M_0)\cdot 10^{-6}+a]$$
$$=[(0.072h+2ge^{-ch})\cdot 10^{-6}+a].$$

The values of g, c, and a necessary to fit the linear exponential to M curves were determined. However, the value of a depends on g and c and may be determined from them by reference 7. In other words, a given shape of M curve must have a certain positioning with respect to the $(y+a)$ axis. Since the positioning is known from the measured data, a comparison may be made between the measured and theoretical position of the $(y+a)$ curve. Such a comparison is shown in Table III.

XIII. CHECK OF ABSOLUTE SIGNAL STRENGTH

The calculation of M curves from radio data requires a knowledge of the relative and not the absolute signal strengths. However, from the shape of the M curves determined from radio data, the signal strength relative to the free-space value may be determined. Using the measured signal strength, a theoretical value for the free-space signal was found for each case by the method of reference 5. Since the gains of the transmitting and receiving system were known as well as the power of the transmitter, a free-space value was established directly. The two values of the free-space signal are compared in Table IV.

From Table IV, it is noticed that the deviation between the theoretical and measured free-space signal strength is a function of the attenuation factor. This relationship is shown graphically in Fig. 9. For strong ducts in which the attenuation factor is low, the measured signal is considerably weaker than the theoretical value. However, for weaker ducts the measured signal strength approaches the theoretical value. It is thought that this deviation may be due to scattering by discontinuities in the atmosphere. For strong ducts, more signal is lost by scattering inside of the duct than is gained from scattering outside of the duct. However, when the signal is weak, the contribution from scattering outside of the duct is significant.

XIV. COMPARISON WITH METEOROLOGICALLY MEASURED M CURVES

As stated in Section III, meteorological data were taken at sites near both ends of the radio path. From the data taken at these sites simultaneously with the radio data, modified index-of-refraction distributions were obtained. For these curves, the measurements made at a given height at the two ends were averaged to obtain the value for that height. The M curves determined from radio data are compared with meteorological data in Fig. 10. From this figure, it is seen that general agreement exists between the M curves determined from radio data and meteorological data. However, it is noted that the meteorologically measured points are badly scattered. It is felt that it is difficult to determine the average M curve over a distance from the measurements at a single site since the meteorological parameters were fluctuating considerably. It seems, therefore, that the M curves may be obtained with a higher degree of accuracy from the radio data than from temperature and humidity measurements.

XV. CONCLUSIONS

It appears from this paper that for the conditions of the experiment, the modified index-of-refraction values as a function of height may be satisfactorily determined from signal-strength and phase-difference measurements. However, for very short distances, the method will fail since a single mode will not account for the signal distributions. For greater distances or very weak signals, the method may fail when an appreciable part of the signal is the result of scattering from elevated discontinuities in the index of refraction.

The Effective Permeability of an Array of Thin Conducting Disks

GERALD ESTRIN*
University of Wisconsin, Madison, Wisconsin
(Received December 5, 1949)

A three-dimensional array of thin conducting disks has found special application as an artificial refracting medium at microwave frequencies. Treatment of these refractive properties at oblique incidence requires specification of the dielectric and permeability coefficients of the medium.

When an alternating magnetic field is parallel to the disk faces, the field is undisturbed and the relative permeability coefficient is unity. When the alternating magnetic field is normal to the disk faces, circulating currents are induced on them. The boundary value problem of determining the current distribution on a single perfectly conducting disk is carried out in detail for the case where the disk diameter is small compared to the wave-length. This current distribution is found to be representable by a magnetic dipole. If the disks in an array are far enough apart to neglect interaction, a simple summation of the dipole moments shows the array to have a diamagnetic susceptibility in the direction normal to the disk faces.

Combining this result with an expression for the dielectric coefficient, which was developed earlier by Kock, the constants of the anisotropic array are completely specified.

INTRODUCTION

THE first part of this paper evaluates the steady state circulating current distribution induced in a thin conducting disk by a perpendicular time varying magnetic field whose wave-length is large compared to the disk diameter. The disk is treated as the limiting case of an oblate spheroid. The use of oblate spheroidal coordinates and the definition of a scalar potential function, $V=rE_\varphi$, lead to a Legendre Polynomial solution outside a conducting spheroid and to a general power series solution inside. Imposition of boundary conditions at the surface and far from the obstacle provides a simple relation for the surface current density distribution in the limit of a perfectly conducting disk.

In Part II, a magnetic dipole equivalent to the surface current distribution on the disk, is evaluated by integration of ring-shaped elements. For the case of a three-dimensional array in which the packing of the disks is small enough to neglect interaction, the dipole moment per unit volume leads to an expression for the magnetic susceptibility. The composite medium is observed to be diamagnetic in the direction normal to the disk faces and non-magnetic in directions parallel to them. There is a simple relation between the magnetic and electric susceptibilities of this array. Knowledge of the disk radius and the number per unit volume completely specifies the component constituitive equations which are necessary to determine the effect of the composite medium on an obliquely incident electromagnetic wave. The constituitive coefficients are calculated for two cases; one typical of microwave lens design and the other of a prism constructed at the University of Wisconsin.

PART I

As illustrated in Fig. 1, a conducting oblate spheroid is placed in a time varying magnetic field which is parallel to the axis of rotation of the spheroid. The applied field is given by $\bar{H}=\bar{\imath}H_a=\bar{H}_0\exp(-j\omega t)$. The diameter of the conducting body is assumed small compared to the wave-length and Maxwell's field equations for steady state become:

$$\nabla\times\bar{E}-j\omega\mu\bar{H}=0, \tag{1a}$$

$$\nabla\times\bar{H}-\sigma\bar{E}=0. \tag{1b}$$

Oblate spheroidal coordinates[1,2] provide a natural system for dealing with this problem. Using the notation defined in Appendix I, the component field equations separate into two independent sets. The set that must be solved in this case is given by:

$$\partial/\partial\eta(h_3E_\varphi)-j\omega\mu h_2h_3H_\xi=0, \tag{2a}$$

$$\partial/\partial\xi(h_3E_\varphi)+j\omega\mu h_1h_3H_\eta=0, \tag{2b}$$

$$\partial/\partial\xi(h_2H_\eta)-\partial/\partial\eta(h_1H_\xi)-\sigma h_1h_2E_\varphi=0. \tag{2c}$$

When the field equations in (1) are uncoupled they give

$$\nabla\times\nabla\times\bar{E}=j\omega\sigma\mu\bar{E}. \tag{3}$$

Since there is only a φ component of $\bar{E}$, (3) yields

$$\frac{\partial}{\partial\xi}\left[\frac{h_2}{h_1h_3}\frac{\partial}{\partial\xi}(h_3E_\varphi)\right]+\frac{\partial}{\partial\eta}\left[\frac{h_1}{h_2h_3}\frac{\partial}{\partial\eta}(h_3E_\varphi)\right]=-j\omega\sigma\mu h_1h_2E_\varphi. \tag{4}$$

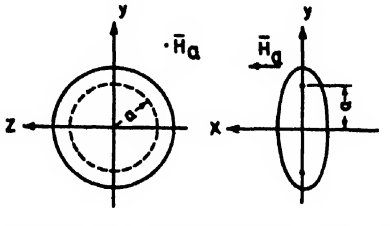


FIG. 1.

* RCA Research Fellow, Electrical Engineering Department.

[1] W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939).
[2] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941).

If a potential function, $V \equiv rE_\varphi$, is now defined and substituted in (2) and (4), the following set of equations results.

$$H_\xi = \frac{-j}{\mu\omega h_2 h_3}\frac{\partial V}{\partial \eta}, \tag{5a}$$

$$H_\eta = \frac{j}{\mu\omega h_1 h_3}\frac{\partial V}{\partial \xi}, \tag{5b}$$

$$\frac{\partial}{\partial \xi}\left[\frac{h_2}{h_1 h_3}\frac{\partial V}{\partial \xi}\right]+\frac{\partial}{\partial \eta}\left[\frac{h_1}{h_2 h_3}\frac{\partial V}{\partial \eta}\right]+j\omega\sigma\mu\frac{h_1 h_2}{h_3}V=0. \tag{5c}$$

Separation of variables in (5c) is carried out in Appendix II yielding identical differential equations in the independent variables, η and $(j\xi)$ of the form,

$$(1-x^2)\partial^2 y/\partial x^2+(k^2x^2+D)y=0, \tag{6}$$

where $k^2=ja^2\omega\sigma\mu$ and a is defined in Appendix I.

Outside the conducting spheroid, $k^2=0$. Appendix III shows that with $D=l(l+1)$, Eq. (6) is then satisfied by a solution containing Legendre Polynomials of the first and second kind.

$$y=(1-x^2)d/dx[P_l(x)+Q_l(x)]. \tag{7}$$

Hence a set of solutions is obtained for the potential function, V, in the region outside the conducting spheroid.

$$V_0=\sum_{l=0}^{\infty} j(1-\eta^2)(1+\xi^2)\left[A_l\frac{d}{d\eta}P_l(\eta)+B_l\frac{d}{d\eta}Q_l(\eta)\right] \times\left[C_l\frac{d}{d\xi}P_l(j\xi)+D_l\frac{d}{d\xi}Q_l(j\xi)\right]. \tag{8}$$

The set of solutions, (8), must now be restricted by imposing the physical conditions of the problem.

(a) The field vectors are bounded. Therefore V_0 must be bounded for finite values of the coordinates. Since $Q_l(\pm 1)$ is unbounded, $B_l\equiv 0$ for all l.

(b) $H_\xi^2+H_\eta^2$ must reduce to the external field H_a^2 when $\xi\to\infty$. As shown in Appendix IV, this condition requires that

$$V_0\underset{\xi\to\infty}{\longrightarrow} j\omega\mu a^2(1-\eta^2)(1+\xi^2)H_a/2.$$

Since

$$Q_l(j\xi)\underset{\xi\to\infty}{\longrightarrow}0 \quad\text{and}\quad P_l(j\xi)\underset{\xi\to\infty}{\longrightarrow}\infty,$$

$$\sum_{l=0}^{\infty}\left[A_l\frac{d}{d\eta}P_l(\eta)\right]\left[C_l\frac{d}{d\xi}P_l(j\xi)\right]$$

may enter into the solution only as a constant. Hence $C_l\equiv 0$ for $l>1$.

(c) Symmetry considerations demand that $V(\xi,+|\eta|)=V(\xi,-|\eta|)$. Hence the solution may contain only even powers of η and $A_l\equiv 0$ for $l=2n$, where $n=0, 1, 2, \cdots$.

Thus the solution for V outside the spheroid reduces to

$$V_0=(1-\eta^2)(1+\xi^2)\left[\frac{j\omega\mu a^2}{2}H_a + \sum_{n=0}^{\infty}F_{2n+1}\frac{d}{d\eta}P_{2n+1}(\eta)\frac{d}{d\xi}Q_{2n+1}(j\xi)\right]. \tag{9}$$

Inside the conducting spheroid, the differential equation to be solved is of the form,

$$(1-x^2)d^2y/dx^2+[l(l+1)+k^2x^2]y=0. \tag{10}$$

A series solution in positive powers of x results in the recursion formula,

$$C_{s+2}=\frac{s(s+1)-l(l+1)}{(s+2)(s+1)}C_s-\frac{k^2}{(s+2)(s+1)}C_{s-2}. \tag{11}$$

The continuity of V at the surface requires that the solution contain only even powers of the independent variable. Hence

$$V_i=\sum_{n=0}^{\infty}G_{2n+1}\sum_s C_{2s}\eta^{2s}\sum_p C_{2p}(j\xi)^{2p}, \tag{12}$$

where the coefficients of (12) and (9) must be matched at the surface.[3]

CURRENT DISTRIBUTION ON A PERFECTLY CONDUCTING DISK

For a perfectly conducting spheroid, $V_i=0$. V must be continuous at the surface, $\xi=\xi_0$, for arbitrary values of η. This requires that

$$-j\mu\omega a^2H_a/2 = \left[\sum_{n=0}^{\infty}F_{2n+1}\frac{d}{d\eta}P_{2n+1}(\eta)\frac{d}{d\xi}Q_{2n+1}(j\xi)\right]_{\xi=\xi_0}. \tag{13}$$

This relation can hold only for $n=0$. Hence $F_{2n+1}\equiv 0$ for $n\neq 0$ giving

$$-j\mu\omega a^2H_a/2=F_1[\cot^{-1}\xi-\xi/(1+\xi^2)]_{\xi=\xi_0}. \tag{14}$$

The spheroid reduces to a thin disk as $\xi\to 0$. In that case (14) yields $F_1=-j\mu\omega a^2H_a/\pi$ and

$$V_0=j(1-\eta^2)(1+\xi^2)\omega\mu a^2H_a \times\{1-2/\pi[\cot^{-1}\xi-\xi/(1+\xi^2)]\}/2. \tag{15}$$

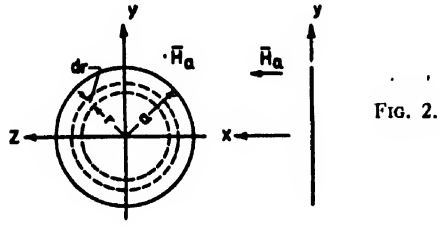


FIG. 2.

[3] Extensive treatment of this type of solution is given by R. C. Maclaurin, Trans. Camb. Phil. Soc. 17, 41-108 (1899).

The surface current density distribution is evaluated from the discontinuity in the tangential component of $\bar{H}$ at the double layer obtained when $\xi \to 0$. Using the relation (5b), the surface current density, i_φ, is

$$i_\varphi = 2[H_\eta|_{\xi=0}] = \frac{2j}{\mu\omega h_1 h_3}\frac{\partial V_0}{\partial \xi}\bigg|_{\xi=0}. \tag{16}$$

When the values of the metric coefficients are substituted, (16) reduces to

$$-i_\varphi = [4a(1-\eta^2)H_a/\pi\eta r]_{\xi=0}. \tag{17}$$

However $(1-\eta^2)/\eta \xrightarrow[\xi\to 0]{} r^2/a(a^2-r^2)^{\frac{1}{2}}$, resulting in the final simple expression,[4]

$$-i_\varphi = 4H_a r/\pi(a^2-r^2)^{\frac{1}{2}}. \tag{18}$$

PART II

The circulating current, I, flowing in the ring element of Fig. 2 is given by

$$I = i_\varphi dr = -4H_a r dr/\pi(a^2-r^2)^{\frac{1}{2}}. \tag{19}$$

The effect of this circulating current may be represented by an elementary magnetic dipole, defined by

$$dm = \pi r^2 I = -4H_a r^3 dr(a^2-r^2)^{-\frac{1}{2}}. \tag{20}$$

Integration over the face of the disk [Dwight 323.01] then provides an expression for a magnetic dipole equivalent to the entire current distribution.

$$m = -(8/3)a^3 H_a. \tag{21}$$

A three dimensional array of thin conducting disks is illustrated in Fig. 3. If the number of disks per unit volume is small enough to neglect interaction between them, the magnetization of the composite medium is given simply by

$$M = Nm. \tag{22}$$

N is the number of disks per unit volume and m is given by (21). Since the ratio of the magnetization, M, to the applied field, H_a, is defined as the magnetic susceptibility x_m,

$$x_m = M/H_a = -8Na^3/3. \tag{23}$$

The negative susceptibility of (23) is a characteristic of diamagnetic media. The relative permeability is then evaluated from the relation,

$$k_m = 1 + x_m = 1 - 8Na^3/3. \tag{24}$$

The positive definite form of the expression for the energy stored in the magnetic field does not permit k_m to be negative. This limitation is implicit in the assumption that N is small.

Up to this point the obvious anisotropy of the composite medium has not been mentioned. For the cartesian coordinates shown in Fig. 3, (24) holds for an applied magnetic field in the x direction, normal to the disk faces. In directions parallel to the disk faces, the thin disks do not disturb the applied field and $k_{my} = k_{mz} = 1$. Thus the constituitive equations for the magnetic vectors take the form,

$$\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} = \mu_0 \begin{bmatrix} 1-8Na^3/3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix}. \tag{25}$$

The electric susceptibility, x_e, for this type of array has previously been evaluated for the case of an external electric field parallel to the disk faces,[5] yielding

$$x_e = 16Na^3/3. \tag{26}$$

When the anisotropy is considered, the constituitive equations for the electric vectors become

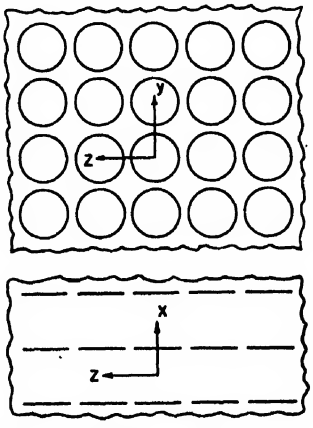


FIG. 3.

$$\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+16Na^3/3 & 0 \\ 0 & 0 & 1+16Na^3/3 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}. \tag{27}$$

The relations in (25) and (27) now completely specify the characteristics of the composite medium, which was originally developed for use in focusing microwaves. Past analyses of this medium have assumed normal incidence of a plane wave, thereby neglecting the anisotropic characteristics. These results now permit extension to the case of oblique incidence.

A typical value for the relative dielectric constant in arrays designed for microwave lenses is $k_e = 2.25$. The constituitive equations for this case then become:

$$\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2.25 & 0 \\ 0 & 0 & 2.25 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}. \tag{28a}$$

$$\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} = \mu_0 \begin{bmatrix} 0.375 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix}. \tag{28b}$$

An experimental prism, constructed at the University of

[4] This result differs by a factor of 2 from that reported by M. Marc Jouguet in an abstract in Comptes Rendus 216, 523 April, 1943. The discrepancy is apparently due to his neglect of the double layer of surface current on the disk.

[5] W. E. Kock, "Metallic delay lenses," Bell Sys. Tech. J. 27, 58 (January, 1948).

Wisconsin displays the following characteristics:

$$\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \epsilon_0 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1.21 & 0 \\ 0 & 0 & 1.21 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}. \quad (29a)$$

$$\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} = \mu_0 \begin{bmatrix} 0.895 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix}. \quad (29b)$$

ACKNOWLEDGMENT

The author would like to acknowledge the support given to this research project by the Radio Corporation of America since July, 1949 and by the Wisconsin Alumni Research Foundation from February, 1948 to June, 1949. In addition the author is indebted to the guidance of Professor V. C. Rideout of the University of Wisconsin Electrical Engineering Department.

APPENDIX I

Oblate spheroidal coordinates are defined by the following relations:

$$y = r\cos\varphi; \quad z = r\sin\varphi, \quad (1a)$$

$$\frac{x^2}{a^2\xi^2} + \frac{r^2}{a^2(1+\xi^2)} = 1 \quad (0 \leq \xi \leq \infty), \quad (1b)$$

$$\frac{x^2}{-a^2\eta^2} + \frac{r^2}{a^2(1-\eta^2)} = 1 \quad (-1 \leq \eta \leq +1), \quad (1c)$$

ξ = constant describes a family of confocal oblate spheroids where (2a) is the distance between the foci in any plane through the x-axis. η = constant describes a family of confocal hyperboloids of one sheet, orthogonal to the spheroids. φ measures the azimuth angle in the meridian plane.

An element of arc length in this system is given by:

$$\begin{aligned} (ds)^2 &= h_1^2(d\xi)^2 + h_2^2(d\eta)^2 + h_3^2(d\varphi)^2, \\ h_1 &= a(\xi^2+\eta^2)^{\frac{1}{2}}(1+\xi^2)^{-\frac{1}{2}}, \\ h_2 &= a(\xi^2+\eta^2)^{\frac{1}{2}}(1-\eta^2)^{-\frac{1}{2}}, \\ h_3 &= r = a[(1+\xi^2)(1-\eta^2)]^{\frac{1}{2}}. \end{aligned} \quad (2)$$

If the field vectors have no φ dependence, the component field equations are given by (2) of Part I along with the following:

$$\partial(h_3H_\varphi)/\partial\eta - h_2h_3\sigma E_\xi = 0, \quad (3a)$$

$$\partial(h_3H_\varphi)/\partial\xi + h_1h_3\sigma E_\eta = 0, \quad (3b)$$

$$\partial(h_2E_\eta)/\partial\xi - \partial(h_1E_\xi)/\partial\eta - j\omega\mu h_1h_2H_\varphi = 0. \quad (3c)$$

APPENDIX II

In (5c) of Part I, substitution of the metric coefficients gives

$$\frac{\partial}{\partial\xi}\left(\frac{1}{a(1-\eta^2)}\frac{\partial V}{\partial\xi}\right) + \frac{\partial}{\partial\eta}\left(\frac{1}{a(1+\xi^2)}\frac{\partial V}{\partial\eta}\right) + j\omega\sigma\mu a\left(\frac{\xi^2+\eta^2}{(1+\xi^2)(1-\eta^2)}\right)V = 0. \quad (1)$$

Let $V = V_1(\xi)V_2(\eta)$ and multiply through by $a(1-\eta^2)(1+\xi^2)/V_1V_2$.

$$\frac{(1+\xi^2)}{V_1}\frac{d^2V_1}{d\xi^2} + \frac{(1-\eta^2)}{V_2}\frac{d^2V_2}{d\eta^2} + ja^2\omega\mu\sigma(\xi^2+\eta^2) = 0. \quad (2)$$

Introducing the separation constant D, and making the substitutions, $\xi' = j\xi$ and $k^2 = j\omega\sigma\mu a^2$ gives

$$(1-\eta^2)d^2V_2/d\eta^2 + (k^2\eta^2+D)V_2 = 0, \quad (3a)$$

$$(1-\xi'^2)d^2V_1/d\xi'^2 + (k^2\xi'^2+D)V_1 = 0. \quad (3b)$$

APPENDIX III

Outside the spheroid,

$$(1-x^2)d^2y/dx^2 + Dy = 0. \quad (1)$$

Consider the equation satisfied by the Legendre Polynomials, $P_l(x)$.

$$(1-x^2)d^2y/dx^2 - 2xdy/dx + l(l+1)y = 0 \quad (2)$$

Differentiate (2) with respect to x and substitute $y = P_l(x)$.

$$(1-x^2)d^2/dx^2[d/dxP_l(x)] - 4xd/dx[d/dxP_l(x)] + [l(l+1)-2][d/dxP_l(x)] = 0. \quad (3)$$

Try the solution $d[P_l(x)]/dx = (1-x^2)^{-r}y$ in (3) and multiply through by $(1-x^2)^r$. This results in

$$(1-x^2)d^2y/dx^2 + 4x(r-1)dy/dx + \{(1-x^2)^{-1}[4x^2r(r-1)] + l(l+1) + 2(r-1)\}y = 0. \quad (4)$$

Let $r = 1$ in (4) giving

$$(1-x^2)d^2y/dx^2 + l(l+1) = 0. \quad (5)$$

Hence $y = (1-x^2)d[P_l(x)]/dx$ is a solution of (1).

A similar procedure shows that $y = (1-x^2)d[Q_l(x)]/dx$ is also a solution.

APPENDIX IV

From (5a, 5b) of Part I and the relation $\partial V/\partial\eta = (\partial V/\partial r)(\partial r/\partial\eta)$,

$$[H_\xi^2 + H_\eta^2] = -1/\omega^2\mu^2r^2[1/h_2^2(\partial r/\partial\eta)^2 + 1/h_1^2(\partial r/\partial\xi)^2](\partial V/\partial r)^2. \quad (1)$$

Introducing the expressions for the metric coefficients given in Appendix I, (1) reduces to

$$[H_\xi^2 + H_\eta^2] = -1/\omega^2\mu^2r^2(\partial V/\partial r)^2. \quad (2)$$

Hence[6]

$$\partial V/\partial r \xrightarrow[\xi\to\infty]{} j\omega\mu rH_\varphi$$

resulting in

$$V_0 \xrightarrow[\xi\to\infty]{} j\omega\mu a^2H_\varphi/2 = j\omega\mu a^2(1-\eta^2)(1+\xi^2)H_\varphi/2. \quad (3)$$

[6] Physical consideration of the final direction of induced current demands the positive root.

Stress Analysis for Compressible Viscoelastic Materials

W. T. Read, Jr.
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received January 11, 1950)

Mathematical methods of stress analysis are presented for linear, compressible, viscoelastic, or anelastic, materials such as metals at high temperatures or high polymers with small strains. For such materials stress, strain and their time derivatives of all orders are related by linear equations with coefficients which are material constants. Fourier integral methods are used to show that static elasticity solutions can be used to determine the time dependent stresses in viscoelastic bodies with any form of boundary conditions.

If stress and double refraction and their time derivatives are also linearly related, the standard photoelastic techniques can be used to determine the directions and difference in magnitude of the time dependent principal stresses, even though the principal stress axes do not coincide with the polarizing axes and both vary with time. When viscoelastic models are used in photoelastic studies, the time variation of the stress distribution in the model represents a first approximation to the dependence of the stress in the elastic prototype on Poisson's ratio.

INTRODUCTION

MANY efforts have been made to extend the standard methods of stress analysis to non-elastic materials. One approach has been to consider the general linear solid, defined as a material in which the stress-strain behavior can be represented by linear relations involving not only stress and strain but also their time derivatives of all orders. An elastic material represents the special case of a linear material in which the time derivatives can be neglected for the range of frequencies involved. Phenomena which are linear but time dependent have been called "viscoelastic" in the case of high polymers[1] and "anelastic" in the case of metals.[2] It has been shown that both the solutions of problems in elasticity and the standard photoelastic techniques can be applied to a linear incompressible solid when the forces and displacements given on the boundary vary uniformly with time and slowly enough that inertial effects can be neglected.[3,4] The present paper extends these results to the general, compressible, linear solid with any form of boundary conditions. The analysis makes use of the standard Fourier integral and operator methods. For simplicity the derivation considers only isotropic materials, but the results are readily generalized to the anisotropic case.

1. STRESS-STRAIN RELATIONS

In this section we show by use of Fourier integral methods that the time dependent stress-strain relations for the general linear solid can be transformed to static relations between the Fourier transforms of stress and strain, these relations having the same form as the elastic stress-strain relations except that the elastic constants are complex functions of frequency, readily determined from a simple tension test. We shall follow the conventional engineering notation, using σ's and ϵ's for normal components of stress and strain respectively and τ's and γ's for shearing components.

The most general linear isotropic stress-strain relations can be expressed in terms of three linear differential operators H, K and L by

$$H(\partial)/(\partial t)\sigma_x = K(\partial)/(\partial t)e + 2L(\partial)/(\partial t)\epsilon_x, \quad (1.1)$$

$$H(\partial)/(\partial t)\tau_{xy} = L(\partial)/(\partial t)\gamma_{xy} \quad (1.2)$$

where $e = \epsilon_x + \epsilon_y + \epsilon_z$ with four similar relations obtain by rotating subscripts. The operators have the form

$$H(\partial)/(\partial t) = H_0 + H_1(\partial/\partial t) + H_2(\partial^2)/(\partial t^2)\cdots H_n(\partial^n)/(\partial t^n)\cdots \quad (1.3)$$

the coefficients being constants of the material. By operating on (1.1) and (1.2) with $[1/H(\partial)/(\partial t)]$, the relations involve only two linear integro-differential operators and have the same form as in the elastic case, the two operators replacing the two independent elastic constants λ and G.

The number of terms in the operators required to adequately represent the stress-strain relation depends on the material. In general an infinite number will be required for an exact representation. It is customary to represent the solid by a mechanical model consisting of springs and dashpots having constants which determine the constants in the stress-strain relations. However, when many derivatives are required in the stress-strain relations, it is not practical to determine the operators. In the general case a more practical approach is to determine two complex elastic constants $G(i\omega)$ and $\lambda(i\omega)$ which are functions of frequency ω and may be determined from a simple experiment and used to express the stress-strain relations in the following way: We represent the time dependence of $\tau_{xy} = \tau_{xy}(t)$ by a Fourier integral

$$\tau_{xy}(t) = \int_{-\infty}^{\infty} \tau_{xy}(\omega) e^{i\omega t} d\omega, \quad (1.4)$$

[1] T. Alfrey, *Mechanical Behavior of High Polymers* (Interscience Publishers, Inc., New York, 1948).
[2] C. Zener, *Elasticity and Anelasticity of Metals* (University of Chicago Press, Chicago, 1948).
[3] T. Alfrey, "Non-Homogeneous Stress in Viscoelastic Media," Quart. J. App. Math. II, 113 (1944).
[4] R. D. Mindlin, "A mathematical theory of photo-viscoelasticity," J. App. Phys. 20, 206 (1949).

where

$$\bar{\tau}_{xy}(\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tau_{xy}(t)e^{-i\omega t}dt \tag{1.5}$$

is the Fourier transform of $\tau_{xy}(t)$. Expressing γ_{xy} also as a Fourier integral, substituting into (1.2), and considering that the relation must hold for all values of time, we have

$$\bar{\tau}_{xy}(\omega)=G(i\omega)\bar{\gamma}_{xy}(\omega), \tag{1.6}$$

where $\bar{\gamma}_{xy}(\omega)$ is the Fourier transform of $\gamma_{xy}(t)$, and the complex modulus

$$G(i\omega)=\frac{L(\partial)/(\partial t)e^{i\omega t}}{H(\partial)/(\partial t)e^{i\omega t}}=\frac{L(i\omega)}{H(i\omega)}. \tag{1.7}$$

Defining another complex modulus,

$$\lambda(i\omega)=[K(i\omega)]/[H(i\omega)], \tag{1.8}$$

the relation (1.1) is transformed to

$$\bar{\sigma}_x(\omega)=\lambda(i\omega)\bar{e}(\omega)+2G(i\omega)\bar{e}_x(\omega), \tag{1.9}$$

where as before the bars denote Fourier transforms. Thus the time dependent stress-strain relations (1.1) and (1.2) are replaced by static relations between the Fourier transforms of stress and strain. The stress-strain relations can also be stated in terms of the complex Young's modulus and Poisson's ratio $E(i\omega)$ and $\nu(i\omega)$, which are related to $\lambda(i\omega)$ and $G(i\omega)$ by the same relations as in the elastic case. For an experimental determination of the two complex elastic constants, it is convenient to perform a simple tension test in which the tension $\sigma_x(t)$ varies with time in any convenient manner. By measuring the strain $e_x(t)$ along, and $e_y(t)$ at right angles to the tension, and taking the Fourier transforms of the three known functions of time, we can calculate

$$E(i\omega)=[\bar{\sigma}_x(\omega)]/[\bar{e}_x(\omega)]$$

and

$$\nu(i\omega)=-[\bar{e}_y(\omega)]/[\bar{e}_x(\omega)]. \tag{1.10}$$

For an anisotropic material the same results hold except that additional complex elastic constants are required.

2. DIFFERENTIAL EQUATIONS AND BOUNDARY CONDITIONS

The differential equations, to be satisfied are the compatibility equations for the strains and the equilibrium equations, which involve the stresses, body forces and inertial terms. Whether we formulate the problem in terms of stress or displacement, we obtain equations which are the same as in the elastic case except that the elastic constants λ and G are replaced by the corresponding operators

$$\frac{K(\partial)/(\partial t)}{H(\partial)/(\partial t)} \quad \text{and} \quad \frac{L(\partial)/(\partial t)}{H(\partial)/(\partial t)}.$$

These equations can be treated in the same way as the stress-strain relations, all quantities including the body forces being expressed as Fourier integrals. This gives equations for the Fourier transforms which are the same as the static elasticity equations except that the elastic constants are replaced by the complex functions of frequency $\lambda(i\omega)$ and $G(i\omega)$. The boundary conditions on the Fourier transforms are the Fourier transforms of the actual time dependent boundary conditions. Thus every stress analysis problem involving the general linear solid can be reduced to a static elasticity problem and the extensive known solutions of elasticity used. When the Fourier transforms have been obtained, by substituting complex elastic constants and boundary values into the elasticity solution, the actual time dependent solution is obtained from the Fourier transforms, using relations of the form (1.4).

3. SPECIAL CASES AND APPROXIMATIONS

In this section we assume that inertial effects and body forces can be neglected, and that the boundary conditions vary uniformly with time. Let the time dependence of the applied forces and displacements be given by an arbitrary function $F(t)$. If the static elasticity solution for $F(t)=\text{constant}=\text{unity}$ is $S_e[x, y, z, \lambda, G]$, then the solution for the general linear solid is

$$S(x, y, z, t)$$

$$=\int_{-\infty}^{\infty}\bar{F}(\omega)S_e[x, y, z, \lambda(i\omega), G(i\omega)]e^{i\omega t}d\omega, \tag{3.1}$$

where

$$\bar{F}(\omega)=\int_{-\infty}^{\infty}F(t)e^{-i\omega t}dt \tag{3.2}$$

is the Fourier transform of $F(t)$, and S may be any component of stress, strain or displacement. When $S_e[x, y, z, \lambda, G]$ is a rational algebraic function of λ and G, Eq. (3.1) can be transformed to the operator equation

$$S(x, y, z, t)$$

$$=S_e\left[x, y, z, \frac{L(\partial)/(\partial t)}{H(\partial)/(\partial t)}, \frac{K(\partial)/(\partial t)}{H(\partial)/(\partial t)}\right]F(t). \tag{3.3}$$

Thus in problems of this type, which include the greater part of practical cases, the procedure is to replace the elastic constants in the elasticity solution for $F(t)=1$ by the corresponding operators. The elasticity solution in operator form acting on the actual $F(t)$ then gives the time dependent solution for the general linear solid. Since the operators are rational algebraic functions of $\partial/\partial t$, the solution can be expressed as a sum of partial fractions of the form

$$\frac{1}{\beta+(\partial/\partial t)}F(t)=e^{-\beta t}\int_{-\infty}^{t}e^{\beta t'}F(t')dt'. \tag{3.4}$$

In cases where the operators are known this is frequently the most practical method of solution.

When the elasticity solution involves only the ratio of elastic moduli, conveniently expressed by the dimensionless quantity ν, special methods and simple approximate procedures can be used. For example, if loads are suddenly applied and held constant, the initial state of stress will be the same as that in an elastic body with Poisson's ratio equal to the Poisson's ratio of the instantaneous elastic response. Then, if the loads are held constant until a steady state is reached, the state of stress approaches that in an elastic body having the Poisson's ratio characteristic of the steady state. This suggests that a general linear solid is equivalent to an elastic solid with a time variable Poisson's ratio. To show that this is true to a first approximation, we expand the static elasticity solution $S_e[x, y, z, \nu]$, for $F(t)=1$, in a Taylor's series about some average value of ν. Then, using a relation of the form (3.1) and neglecting second and higher order terms in the Taylor's series, we find

$$S[x, y, z, t]=S_e[x, y, z, \nu(t)]F(t), \qquad (3.5)$$

where $\nu(t)$ is determined in terms of $F(t)$ by a simple tension test in which the strain e_x in the direction of tension is $F(t)$, and $\nu(t)=-e_y(t)/F(t)$, $e_y(t)$ being measured at right angles to the tension Equation (3.5) gives a first approximation to the time variation of the solution, and is exact in the special case where $\nu(t)$ = constant. The incompressible case is a special case of this special case, and corresponds to $\nu(t)=\frac{1}{2}$.

Another simple case occurs in problems of plane strain where the boundary conditions are on stress and there is no resultant force on an internal boundary; then the elasticity solution for the stress distribution is independent of the elastic constants, and the stress distribution in the general case is $F(t)$ times the static elasticity solution.

4. OPTICAL METHODS

In this section we consider transparent materials in which the stress-optical relations are of the same form as the stress-strain relations, and can be expressed by linear differential equations with constant coefficients. Although operator methods are used in the derivation, the parallel Fourier integral treatment is evident. It is also known that the standard techniques used in photoelasticity can be employed to determine the directions of the secondary principal axes of stress in the plane of the wave front and the secondary principal stress difference. No restrictions on the boundary conditions are involved.

The stress optical effect involves the relation between the stress and velocity tensors, the latter being simply related to the dielectric constant tensor.* We let V_{xx}, V_{xy}, $\cdots V_{zz}$ be the components of the velocity tensor. Since this tensor is symmetric, it can be represented by an ellipsoid, called the velocity ellipsoid, which provides a graphic means of describing the polarizing axes and wave velocities for plane polarized light. If the xy plane is the plane of the wave front, then the polarizing axes are the axes of the ellipse formed by the trace of the ellipsoid in the xy plane and the wave velocities, V_1 and V_2, are inversely proportional to the squares of the major and minor axes of the ellipse. The standard photoelastic techniques give the difference V_1-V_2 and the angle φ between the polarizing axes and the fixed x and y axes. The difference $V_{xx}-V_{yy}$ is given by the tensor relation

$$V_{xx}-V_{yy}=(V_1-V_2)\cos 2\varphi, \qquad (4.1)$$

where all quantities vary with both time and position in the specimen.

For an isotropic viscoelastic material the most general linear differential equations for the time dependence of the stress-optical effect give

$$X(\sigma_x-\sigma_y)=Y(V_{xx}-V_{yy}), \qquad (4.2)$$

where X and Y are linear differential operators which may have any number of terms. Combining (4.1) and (4.2) gives

$$\sigma_x-\sigma_y=\frac{Y}{X}(V_1-V_2)\cos 2\varphi, \qquad (4.3)$$

from which the stress difference $\sigma_x-\sigma_y$ can be calculated as a function of time for all points, where V_1-V_2 and φ have been measured.

The choice of the fixed axes x and y is arbitrary. The stress difference $\sigma_x'-\sigma_y'$ referred to axes x' and y' at an angle θ to the x and y axes is given by (4.3) with φ replaced by $\varphi-\theta$. Maximizing this with respect to θ gives the secondary principal stress difference in the plane of the wave front, the value of θ giving the direction of the secondary principal axes. Thus, as in photoelasticity, the optical measurements give the directions and difference in magnitude of the secondary principal stresses. In the general viscoelastic case the principal axes of stress are not parallel to the polarizing axes.

In the special cases of Section 3 where the state of stress is known to be the same as the stress in an elastic body, (4.3) takes the simple form

$$\sigma_1-\sigma_2=K(t)(V_1-V_2), \qquad (4.4)$$

where

$$K(t)=\frac{1}{F(t)}\frac{Y}{X}F(t)$$

and σ_1 and σ_2 are the secondary principal stresses and

* The principal axes of the velocity tensor coincide with the principal axes of the dielectric constant tensor and the principal velocities are inversely proportional to the square roots of the principal dielectric constants: G. Joos, *Theoretical Physics* (G. E. Stechert and Company, New York, 1934), p. 349.

act in the directions of the polarizing axes. A simple tension experiment gives $K(t)$.

When viscoelastic materials are used for models in optical stress studies, and the Poisson's ratio has been measured for the instantaneous elastic deformation under constant load and also for the final steady state deformation, then the photoelastic analysis gives the stress distribution in the elastic prototype for two values of Poisson's ratio. As was discussed in the last section, the intermediate stress distributions are the first approximations to the stresses in an elastic prototype with a time variable Poisson's ratio. The viscoelastic model therefore gives information about the dependence of the stress distribution on Poisson's ratio.

Electrolytic-Tank Measurements for Microwave Metallic Delay-Lens Media*

SEYMOUR B. COHN
Sperry Gyroscope Company, Great Neck, New York
(Received December 15, 1949)

In this paper it is shown how the low frequency index of refraction of a metallic delay-lens medium may be calculated from electrolytic-tank measurements on individual metallic elements of the medium. The proximity between adjacent elements is taken exactly into account for low frequencies. The test apparatus is described, and measured results are presented for three kinds of delay-lens structures.

I. INTRODUCTION

IT has been shown by W. E. Kock that an array of discrete metallic obstacles in free space (Fig. 1) has an index of refraction greater than unity, and hence is a suitable lens medium for microwaves.[1] In his article Kock gives approximate index-of-refraction formulas, surface-shape relations, and practical methods of construction.

Kock's analysis for thin-strip and thin-circular-disk obstacles assumes the dimensions of the obstacles to be small compared to their spacings. Because of this assumption, substantial errors exist in practical delay-lens media for which rather close spacings are required in order to obtain a reasonably high index of refraction. In the case of the metallic-strip structure, the writer has presented accurate theoretical relations and graphs for the index of refraction and image admittance with proximity between the strips and the effect of frequency taken into account.[2] For other obstacle shapes an exact theoretical solution appears to be exceedingly difficult, and therefore experimental methods of acquiring index-of-refraction data are necessary. One possible method calls for direct index-of-refraction measurements in the microwave range on delay-medium samples. Another method, which is described in this paper, utilizes indirect measurement of the capacitance of individual obstacles with the aid of an electrolytic-tank analogy. Since promixity is taken fully into account, the index of refraction computed with the electrolytic-tank analogy is exact if the wave-length is much greater than the size and spacing of the obstacles.

The writer has shown how the problem of a metal-strip delay-lens medium may be reduced to that of a parallel-plane transmission line containing uniformly spaced obstacles, and has given formulas by which the parameters of the medium may be calculated from the equivalent circuit of the transmission line.[2] The same

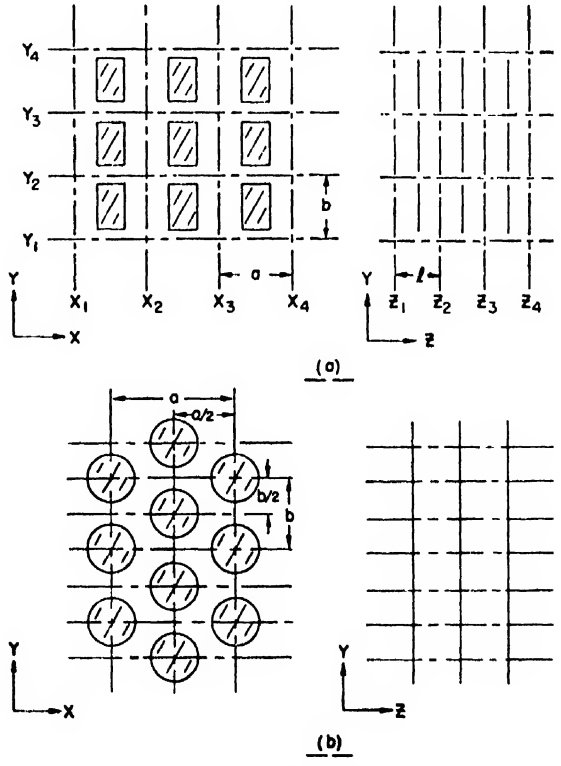


FIG. 1. Delay-lens media.

* The experimental work described in this paper has been performed as a part of a program of research under Contract No. W36-0.39-sc-38250 for the Evans Signal Laboratory, U. S. Army Signal Corps.

[1] W. E. Kock, "Metallic delay lenses," Bell Sys. Tech. J. 27, 58 (1948).

[2] S. B. Cohn, "Analysis of the metallic-strip delay structure for microwave lenses," J. App. Phys. 20, 257 (1949).

reduction may be made with many other kinds of delay-lens media (for example, those of Fig. 1) and the same formulas apply. The purpose of the electrolytic-tank measurements is to provide quantitative values of the shunt capacitance introduced by each obstacle in the equivalent circuit.

II. THE EQUIVALENT CIRCUIT

The procedure by which an equivalent transmission line may be obtained for the two configurations of Fig. 1 will now be explained. Assume a plane wave in

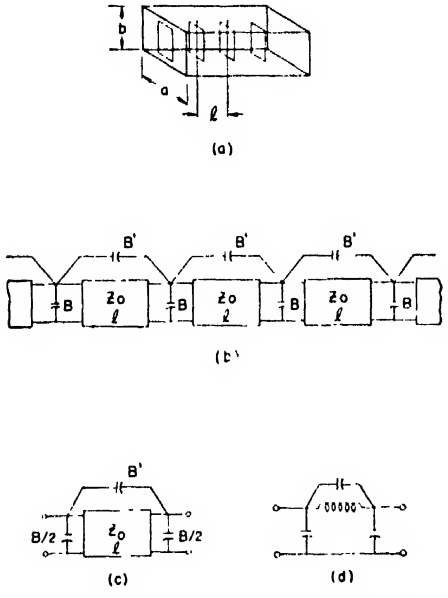


FIG. 2. Equivalent circuit of obstacles in a parallel-plane line.

the medium traveling in the z direction with the average electrical polarization in the y direction. Then with no disturbance to the electromagnetic field, thin "electric" walls may be inserted horizontally in planes y_1, y_2, $y_3 \cdots$ and thin "magnetic" walls vertically in planes x_1, x_2, $x_3 \cdots$ (an electric wall is a conducting sheet, while a magnetic wall is a hypothetical surface on which only tangential-electric and normal-magnetic field components can exist. Such a surface does not occur in nature, but is a useful tool for theoretical analysis). In this way a single channel through the medium such as that in Fig. 2a has been isolated. Since the top and bottom surfaces are conductors while the sides are nonconductors, the channel is in reality a parallel-plane transmission line shunted at uniform spacings by capacitive metallic obstacles. The wave velocity and the normalized image impedance of the transmission line is the same as that of the corresponding lens medium.

For the hexagonal distribution shown in Fig. 1b, the same procedure applies for obtaining the equivalent transmission line. Note that electric and magnetic planes may be passed through the center of each circular obstacle without altering the field, and hence a parallel-plane line having the cross-sectional dimensions $a/2$ and $b/2$ may be used.

The equivalent circuit for the basic transmission line of the metal-strip delay medium has been found in reference 2 to be that of Fig. 2b. The shunt capacitive susceptances represent the shunting effect of the obstacles in the transmission line, while the bridging capacitive susceptances B' represent a proximity effect between adjacent obstacles in the line. Proximity effects between nonadjacent obstacles are neglected, since they are not believed to be of appreciable magnitude in delay-media computations. A second effect of proximity is a reduction in the magnitude of B. Experience with the metal-strip medium has shown that B' is negligible when the obstacle spacing l is large enough to make B constant. This condition occurs for l/b greater than approximately 0.75.

The boxes in Fig. 2b represent sections of transmission line of length l and characteristic impedance Z_0. A single section of the line is shown in Fig. 2c. When the Pi equivalent circuit of the length l of uniform line is substituted for the box and combined with the obstacle susceptances, one obtains the circuit shown in Fig. 2d, which may be recognized as an m-derived low pass filter section. Consequently, the delay lens may be expected to transmit waves freely up to a definite cut-off frequency and to have an opaque band above that. For lens applications only the band below this cut-off frequency is useful, and all discussion in this paper will be limited to this range. Even within this range, however, the capacitances and inductances of the equivalent circuit are not constants, but vary somewhat with frequency.

Note in Fig. 2d that the bridging susceptance B' is shunted by an inductive element. At low frequencies B' is almost zero while the inductive susceptance is very large, and therefore it may be seen that B' can have no effect on the low frequency behavior of the delay medium. Consequently, the electrolytic-tank method can give precise values of the index of refraction at low frequencies even though it is incapable of measuring B'.

The formulas for the index of refraction and image

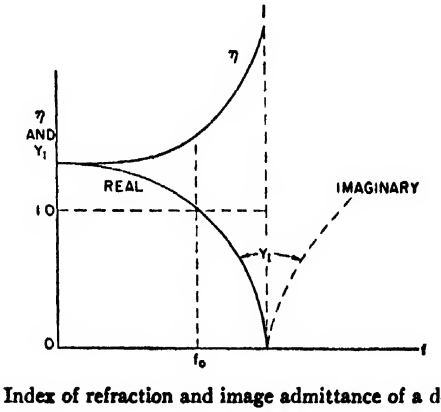


FIG. 3. Index of refraction and image admittance of a delay lens.

SPACE-TANK ANALOGUE (MKS UNITS)	
SPACE	TANK
E	E
V	V
D	i
$Q=\int D\cdot ds$	$I=\int i\,ds$
C=Q/V	G=I/V

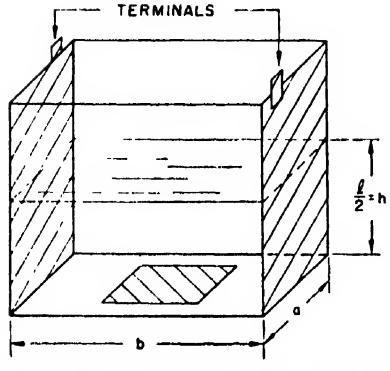


FIG. 4. Electrolytic analog for one obstacle.

admittance for the case of B' negligible are:[2]

$$\eta=\frac{\lambda}{2\pi l}\cos^{-1}\left(\cos\frac{2\pi l}{\lambda}-\frac{B}{2Y_0}\sin\frac{2\pi l}{\lambda}\right), \tag{1}$$

$$y_I=\frac{Y_I}{Y_0}=\left[1+(B/Y_0)\cot\frac{2\pi l}{\lambda}-(B/2Y_0)^2\right]^{\frac{1}{2}}, \tag{2}$$

where η is the index of refraction, λ the wave-length in free space, Y_0 the characteristic admittance of the parallel-plane transmission line, Y_I the image admittance of the line loaded by the obstacles, y_I the normalized image admittance, and l the spacing shown in Fig. 1. Equations (1) and (2) are correct only for a plane wave traveling normal to the obstacle planes. The surfaces of the lens should therefore be shaped in a given application so as to obtain this condition inside the lens. For this purpose, the index of refraction given by Eq. (1) may be correctly used with Snell's law of refraction. This was done by Kock in deriving his lens-surface formula for a "pencil-beam" lens.[1]

Figure 3 shows a sketch of η and y_I. Note that at low frequencies these functions are essentially constant and equal to each other. At higher frequencies η bends upward and y_I downward. When B' is neglected they bend less sharply than when it is taken into account. The limiting value of η and y_I at zero frequency is given by

$$\eta_0=y_{I0}=(1+B\lambda/2\pi lY_0)^{\frac{1}{2}}. \tag{3}$$

Since the construction and test of microwave delay-lens models are difficult and expensive, a static method of obtaining the lens parameters is desirable, even though it is accurate only at a low enough frequency or for a large enough l/b ratio so that B' is negligible. Direct measurements on a single obstacle are impossible in free space, since the necessary magnetic-wall boundary conditions are physically nonrealizable. Such measurements are feasible in an electrolytic tank, however, as will be apparent from the following discussion.

III. PRINCIPLE OF THE ELECTROLYTIC-TANK METHOD

If a static potential is impressed between two conductors, the electric field lines will follow the same paths whether the conductors are immersed in free space or in a conducting solution. In Fig. 4 the analogy between various free-space quantities and electrolytic-tank quantities is brought out. The electric field and the potential in space are analogous to the same quantities in the tank, the displacement flux $D=\epsilon E$ in space is analogous to the current density $i=\sigma E$ in the tank, the charge on a conducting surface in space is analogous to the total current flowing through the surface in the tank, and the capacitance between two surfaces in space is analogous to the conductance between surfaces in the tank. In addition, an electric wall in space is the analog of a conducting wall in the tank, while a magnetic wall in space is the analog of a non-conducting wall in the tank. The latter point is of extreme importance, since it shows that the boundary conditions for the equivalent transmission line may be correctly simulated in the tank, thus permitting a precise low frequency determination of the delay-medium parameters with proximity taken completely into account.

It will now be shown that it is only necessary to make measurements on a length $l/2$ of the transmission line in order to determine B correctly. Consider a static potential impressed between the top and bottom surfaces of the line of Fig. 2a. Then without disturbing

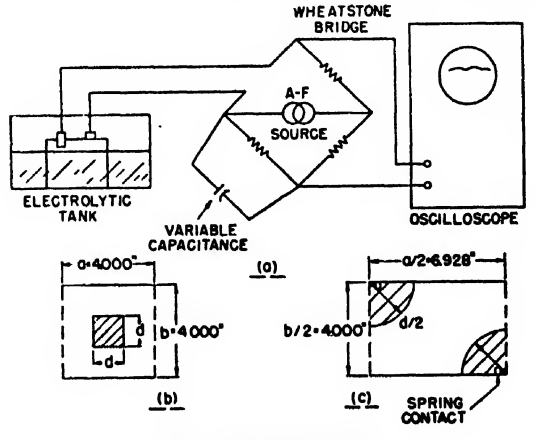


FIG. 5. Test equipment.

the electric field, magnetic walls may be inserted across the line in the obstacle planes and also halfway between the obstacles. The cell in Fig. 4 shows the electrolytic analog of the space between two magnetic walls. The obstacle and the two vertical walls of length a are conductors. The two vertical walls of length b, the surface of the solution, and the bottom of the cell (except for the obstacle) are all non-conductors, simulating the magnetic walls of their free-space counterparts. In order to determine the capacitance of the obstacle, the conductance between the metallic sides

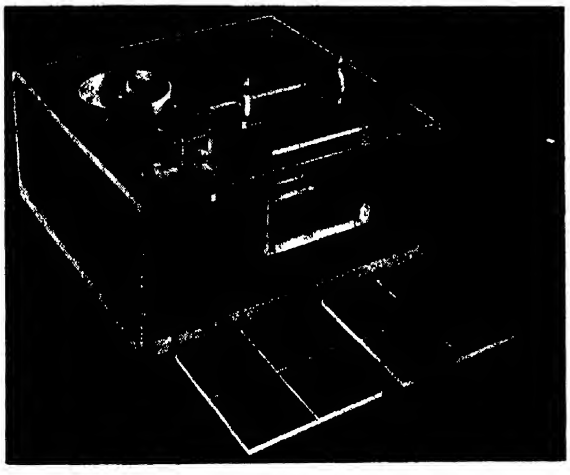


FIG. 6. Photograph of electrolytic-tank apparatus.

of the cell is measured with and without the obstacle present. The apparatus and techniques of the test will now be described.

IV. THE APPARATUS

Figure 5 shows a block diagram of the test equipment. The cell conductance is measured by means of a Wheatstone bridge. The arms of the bridge consist of precision decade resistance boxes having a range of 1 to 111,110 ohms in one-ohm steps. The signal source operates at 1000 cycles and has a balanced ungrounded output. The detector is a high gain oscilloscope. the variable capacitance across one of the bridge arms serves to balance out stray circuit capacitances and a small series reactance which occurs at the contact between electrolyte and metal. The sensitivity of the circuit is such that a change in a bridge arm of one part in five thousand is easily discernible.

Rhodium plating is used on all conducting surfaces in contact with the liquid, since rhodium is highly inactive chemically and electrolytically. The electrolyte is a dilute solution of potassium chloride in distilled water. The concentration was chosen so that measured resistances would lie in the range of about 1500 to 9000 ohms. These resistances are too low for stray shunt capacitances to cause significant error and too high for circuit resistance and inductance to affect the accuracy.

A photograph of the tank apparatus appears in Fig. 6. The tank is made of Lucite and has internal dimensions of 12×12×5 in. The cell, which has dimensions of 4×4×3 in., has four vertical Lucite walls and no top or bottom. One opposite pair of walls is rhodium plated on the inside, and terminals connect to each of these surfaces. The obstacles are cut from thin copper sheets 0.0015 in. thick and are rhodium plated and cemented onto Lucite plates. Grooves are cut in the Lucite on the opposite side so that the plates will flatten readily with pressure when in use. The coaxial cylinders shown in the photograph are used to calibrate the conductivity of the solution. Although it is theoretically possible to obtain this calibration with the cell with the obstacle removed, it has been found that small but not negligible leakage currents escape from the interior of the cell through the cracks between the cell and the bottom plate, thus adding an unknown but constant conductance to the readings with and without the obstacle in place. The coaxial cylinders are by nature self-shielding, and leakage currents cannot exist. Hence, they may be expected to provide an accurate conductivity calibration. A depth gauge is also visible in Fig. 6. By means of this, the height of the electrolyte may be measured to an accuracy of about 0.002 in.

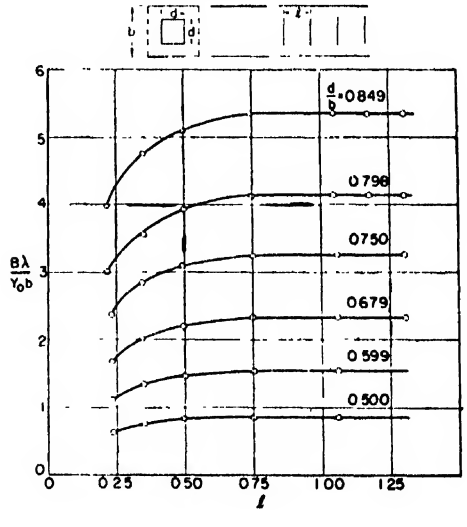


FIG. 7. Susceptance of thin square obstacles.

In performing a test, resistance measurements are made on the coaxial cylinders and on the cell with and without an obstacle in position. For the measurements without the obstacle, a blank Lucite plate having the same thickness as the obstacle plate is used in order to maintain a constant solution height. The susceptance of the obstacle may then be calculated from the following formula which is derived in the Appendix.

$$\frac{B\lambda}{Y_0 b} = \frac{4\pi h}{a}\left(\frac{376.6}{Z_{0\,\mathrm{coax}}}\right)\frac{R_3(R_2-R_1)}{R_1R_2}, \qquad (4)$$

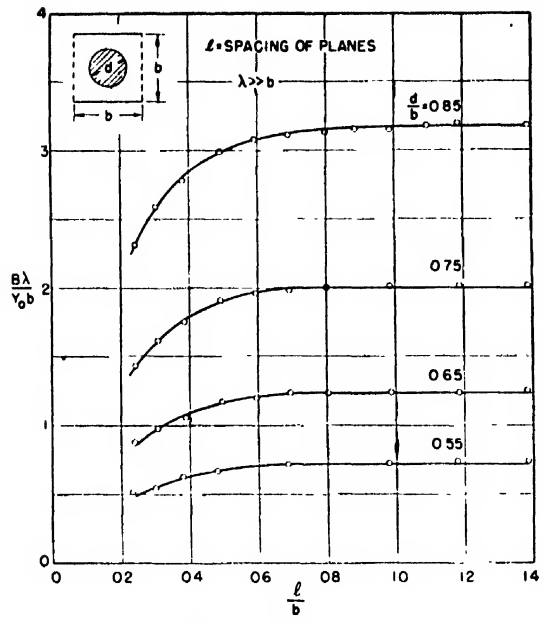


FIG. 8. Susceptance of thin circular obstacles, square distribution.

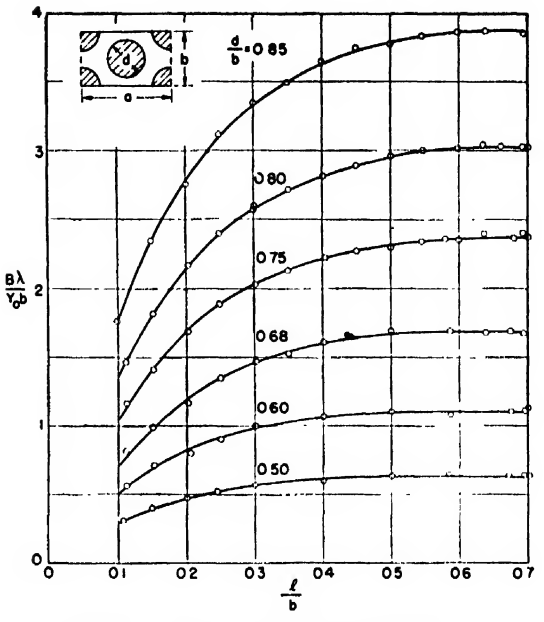


FIG. 9. Susceptance of thin circular obstacles, hexagonal distribution.

where B = shunt susceptance of obstacle ($=\omega C$); $Y_0 = 1/Z_0$ = characteristic admittance of parallel-plane line ($=a/376.6b$); λ = wave-length; b = distance between metal plates of line; a = width of metal plates; h = height of water ($=l/2$); $Z_{0\,coax}$ = characteristic impedance of coaxial cylinders ($=60 \ln r_2/r_1$); R_3 = resistance of coaxial cylinders; R_2 = resistance of cell without obstacle; R_1 = resistance of cell with obstacle.

Three kinds of configurations were measured, square obstacles having square distribution (Fig. 1a), circular obstacles having square distribution, and circular obstacles having hexagonal distribution (Fig. 1b). A top view of the cell for the first two is shown in Fig. 5b (with a square obstacle in place). A top view of the cell for the hexagonal distribution is shown in Fig. 5c. In this case it is necessary to have a low resistance contact between each obstacle quadrant and the adjacent conducting wall. This is obtained by means of a small leaf spring soldered to the wall. Since the spring is small and in a region of negligible current flow, its presence causes no appreciable error.

V. EXPERIMENTAL DATA

The measured shunt-susceptance data for the three configurations are shown in Figs. 7–9. Also, a table of values taken from accurately drawn curves plotted from the data is given in the Appendix. Notice that for l/b greater than approximately 0.75 the obstacle susceptances are independent of l. In practical delay media, however, it is generally necessary to have l/b less than 0.75 in order to have a large enough index of refraction, and hence the non-constant portions of the graphs are the most useful. The data of Figs. 7, 8, and 9 are valid only if $\lambda \gg b$, since for shorter wave-lengths the obstacle susceptance does not have the frequency dependence of a constant capacitance. Some experimental r-f data now available indicate that the frequency error is very small if $b/\lambda \leq 0.25$.

The following theoretical formula gives the low frequency susceptance in a parallel-plane line of one or more thin circular disks having a diameter much smaller than the spacings to surrounding walls and having $l/b > 0.75$.

$$B/Y_0 = 4\pi n d^3/3\lambda ab. \tag{5}$$

This formula may be derived very easily from a relation given by W. E. Kock for the average dielectric constant of a medium containing a uniform distribution of circular disks (Eq. (22), reference 1). The symbol n in Eq. (5) is the total number of obstacles in a cross section of the equivalent transmission line of sides a and b. It is equal to one for the square distribution and two for the hexagonal distribution. Equation (5) is plotted in Fig. 10 along with the experimental data for circular disks ($l/b > 0.75$). The agreement between experiment and simple theory is very good for d/b less than 0.5. This result checks the experimental method and also shows that Kock's low frequency formula for index of refraction is accurate provided that d/b is less than 0.5 and l/b greater than 0.75.

Another check between experiment and theory was obtained with a metal-strip obstacle which extended the full width of the cell and which had a ratio of 0.6 between the width of the strip and the spacing of the conducting plates. In this case the measured data are within one percent of the exact formula.

In Figs. 11–13 are plotted low frequency index-of-refraction curves calculated by Eq. (3) from the ex-

perimental curves of Figs. 7, 8, and 9. Notice that for given obstacle spacings the square obstacles have the highest index of refraction, the circular obstacles with hexagonal distribution the next highest, and the circular obstacles with square distribution the least. Consequently, the first two structures are likely to be more useful than the third.

The curves of Figs. 11–13 are accurate for b/λ up to approximately 0.1. For b/λ up to approximately 0.2, Eq. (1) may be used. For larger b/λ it is necessary to take account of B', which cannot be determined directly by any static method. The writer believes that a judicious estimate of B' may be made based on the measurements of B, but a discussion of this possibility will be postponed, since experimental verification is thus far lacking.

Although the measurements on the hexagonal distribution were made with the incident electric field parallel to the y axis (Fig. 1b), the index of refraction of a normally incident wave is the same for any polarization. This statement is obvious from the fact that three polarizations of a normal wave exist for which the field distributions are identical (except for 120 degree rotations), and that a normal wave of any other polarization may be represented by the sum of components polarized on two of these three axes. In a similar manner, it may be seen that the index of refraction of the square distributions is independent of the polarization of a normally incident wave. In addition, the normalized susceptance B'/Y_0' of the hexagonal distribution for the electric field parallel to the x axis is identical to B/Y_0 computed from Fig. 9 for the y axis polarization, and the normal susceptance of the square distributions for polarizations of ± 45 degrees from the x or y axis is identical to B/Y_0 computed from Figs. 7 and 8.

The writer wishes to thank his colleagues, H. L. Busching, R. Reisley, and R. C. MacVeety, who assisted in perfecting the measurement apparatus and techniques and in taking the data.

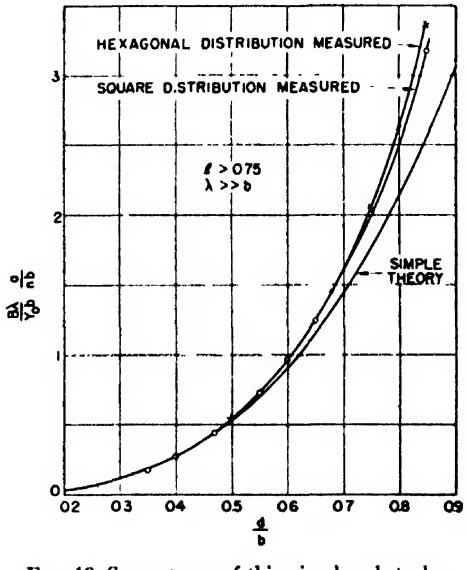


FIG. 10. Susceptance of thin circular obstacles.

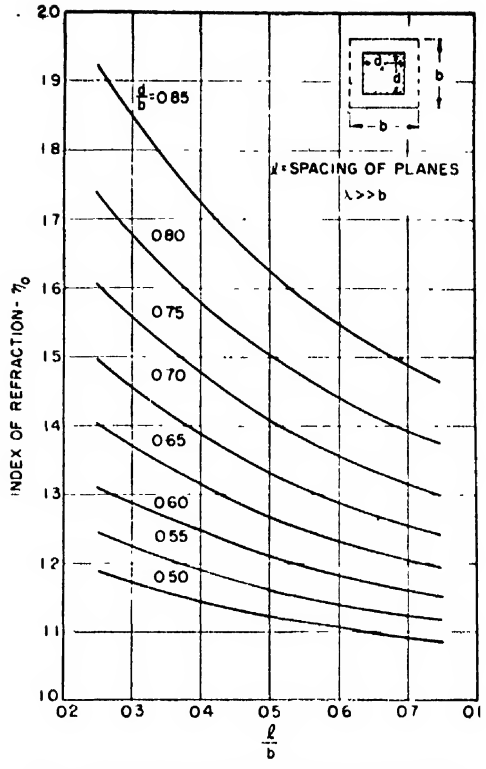


FIG. 11. Index of refraction of square obstacles.

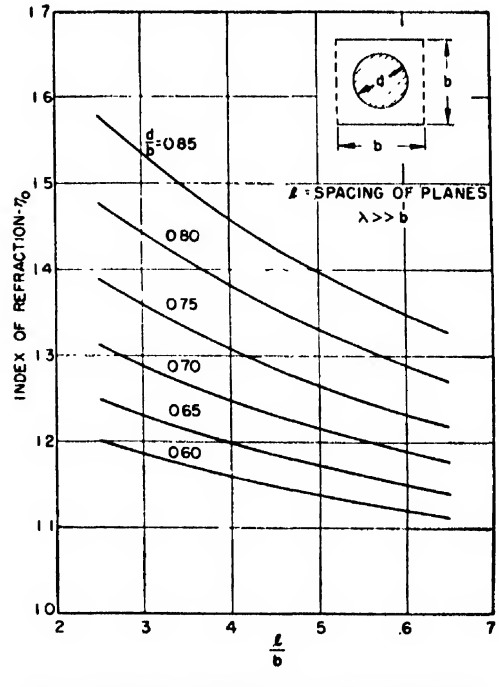


FIG. 12. Index of refraction of circular obstacles, square distribution.

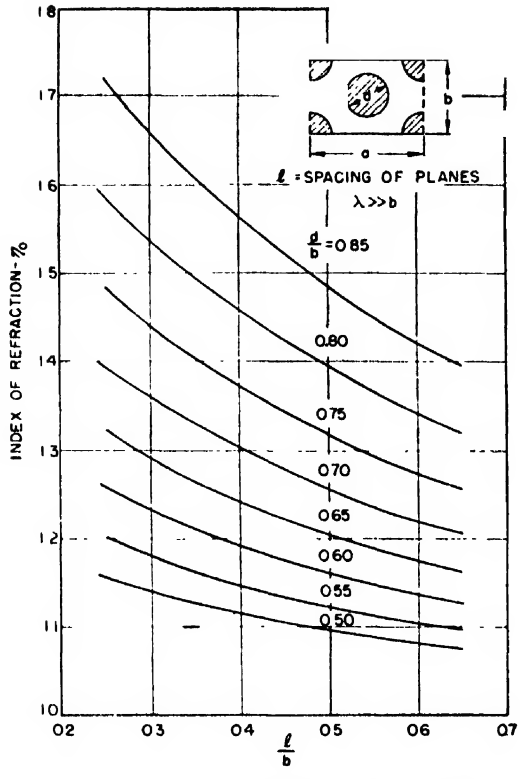


FIG. 13. Index of refraction of circular obstacles, hexagonal distribution.

APPENDIX I

1. Derivation of Eq. (4)

Let the capacitance introduced by the obstacle be C_{obst} and that of the coaxial line C_{coax}. The corresponding conductances in the tank are G_{obst} and G_{coax}. Because of the analogy between capacitance and conductance

$$C_{obst}/C_{coax}=G_{obst}/G_{coax}.$$

The capacitance of the coaxial line is

$$C_{coax}=h/cZ_{0\,coax}\text{ farads},$$

where c is the velocity of propagation in free space. The conductances are obtained from the measured resistances by

$$G_{coax}=1/R_2$$
$$G_{obst}=1/R_1-1/R_2=(R_2-R_1)/(R_1R_2).$$

In the described test method, the obstacle is cemented on a flat plate, and hence only half of the total obstacle capacitance C is measured. Therefore

$$C=2\left[\frac{h}{cZ_{0\,coax}}\right]\frac{R_2(R_2-R_1)}{R_1R_2}.$$

When this is combined with $B=\omega C=2\pi cC/\lambda$ and $Y_0=a/376.6b$ one obtains Eq. (4).

2. Values of $B\lambda/Y_0b$

(a) Square obstacles.

d/b	$l/b=0.25$	$l/b=0.35$	$l/b=0.50$	$l/b=0.65$	$l/b\geq 0.75$
0.50	0.648	0.768	0.830	0.855	0.859
0.60	1.16	1.35	1.475	1.54	1.56
0.70	1.95	2.25	2.45	2.53	2.57
0.75	2.48	2.87	3.11	3.21	3.25
0.80	3.18	3.62	3.97	4.13	4.18
0.85	4.25	4.78	5.15	5.32	5.37

(b) Circular obstacles, square distribution.

d/b	$l/b=0.25$	$l/b=0.35$	$l/b=0.5$	$l/b=0.65$	$l/b\geq 0.8$
0.60	0.68	0.81	0.93	0.97	0.98
0.70	1.13	1.34	1.51	1.58	1.60
0.75	1.45	1.71	1.91	1.98	2.01
0.80	1.85	2.17	2.40	2.50	2.54
0.85	2.35	2.72	3.00	3.11	3.17

(c) Circular obstacles, hexagonal distribution.

d/b	$l/b=0.25$	$l/b=0.35$	$l/b=0.5$	$l/b\geq 0.65$
0.50	0.525	0.595	0.630	0.635
0.60	0.915	1.025	1.095	1.100
0.70	1.48	1.69	1.82	1.86
0.75	1.89	2.13	2.31	2.38
0.80	2.39	2.71	2.96	3.03
0.85	3.09	3.50	3.78	3.87

The Role of Cathode Temperature in the Glow Discharge

HAROLD JACOBS AND JACK MARTIN
Signal Corps Engineering Laboratories, Fort Monmouth, New Jersey
(Received March 24, 1950)

Experiments were conducted to study the role of the temperature of oxide-coated cathodes in the glow discharge of argon, neon, and mixtures of neon and argon, and mercury vapor and argon.

It was found that for tubes without mercury present, in the region of cathode temperature between 300°K and 750°K: (1) The sparking potential falls with increasing cathode temperature. (2) The regulation voltage rises with increasing cathode temperature. For argon and mercury vapor, the regulation voltage falls with increasing cathode temperature. The action here is attributed to metastable states of inert gas. (3) The current density in the glow discharge decreases with increasing cathode temperature. This, too, is attributed to the action of metastable states. (4) The minimum sustaining current in a glow discharge is found to decrease with increasing cathode temperature and rise with increasing pressure. (5) The current density in the glow discharge varies roughly linearly with pressure, or at a slightly higher rate, in contrast with the space-charge derivations which claim a variation with pressure squared. (6) The regulation voltage tends to rise very slightly at lower pressures.

I. INTRODUCTION

COLD cathode behavior in the glow discharge region has received more interest in recent years due to its applications to electronic devices such as thyratrons, voltage regulators, lighting tubes, and switching devices. In the utilization of a glow tube in an electronic circuit a considerable amount of fundamental information must be integrated. For instance, variations of the cathode fall in potential is one type of problem which must be appreciated in designing a voltage regulator. The sparking potential values and relative stability of sparking potentials is important in switching devices and some thyratron control circuits. In the design of glow tubes, the current densities are of importance in order to calculate wattage dissipations at the cathode.

In reference to the stability of electrical performance of glow tubes, there is little or no fundamental information concerning the materials being used in modern tube fabrication techniques such as tubes with oxide-coated and inert gases. There is some general information, however, which is of interest and use in discussing these matters. For instance, the cathode fall in potential has been found to be related to the chemistry of the cathode surface. Bär[1] has shown that a lower work function will generally indicate a lower cathode fall in potential, for a given gas. Loeb[2] points out some early work in which analytic expressions were derived for current density in a glow discharge and cathode fall in potential.

In all of these calculations and in the derivations of most other investigators, the physical processes occurring in the glow are described as follows.

The mechanism of liberation of electrons from the cathode by positive ions or excited atoms is supposed to occur primarily in three ways. (a) Liberation of electrons by positive ion bombardment. (The exact mechanism is not yet understood.) (b) The freeing of electrons by photons produced in the gas. (c) Electron emission due to the effects of metastable atoms releasing energy to the cathode surface.

The electrons liberated from the cathode will, when accelerated, collide with the gas causing ionization and metastable states. These will, in turn, produce more electrons at the cathode. Most of these gas excitation processes are assumed to occur in the region of the cathode glow. For glow tubes where the current density is low and adequate anode area available, the cathode fall in potential is nearly equal to the total voltage across the tube.

In almost all of the reported experimental work with glow discharges and cathode effects, the tubes studied utilized metallic cathodes. In addition, for the case of these metallic surfaces, temperature changes were generally not regarded as important, until the higher temperatures were reached.[3]

II. STATEMENT OF THE PROBLEM

The purpose of the investigation was to describe the action of the glow discharge for the case of an oxide-

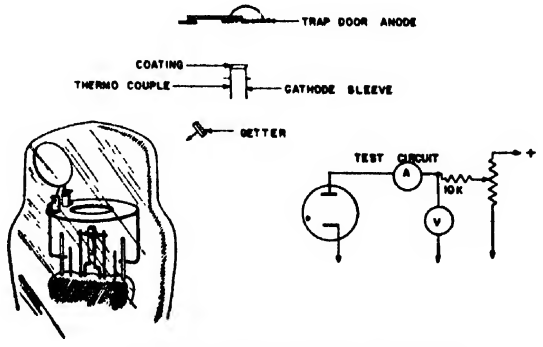


FIG. 1. Experimental tube and test circuit.

[1] Bär, Geiger, and Scheel, *Handbuch der Physik* (1927), Vol. 14, p. 203.
[2] L. B. Loeb, *Fundamental Processes and Electric Discharges in Gases* (John Wiley and Sons, Inc., New York, 1939), pp. 580–585.
[3] See reference 2, p. 607.

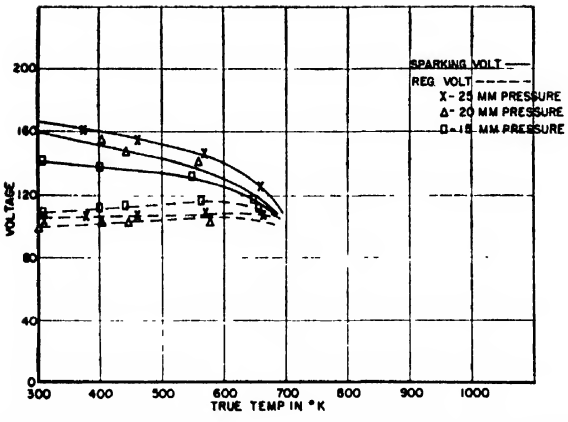


FIG. 2. Sparking potentials and regulation voltages as a function of temperature for neon gas and (BaSrCa)O cathodes.

coated cathode and an inert gas. Specifically, the following factors were studied as a function of cathode temperature and gas pressure: (a) sparking potential, (b) tube voltage drop, (c) current density, and (d) minimum current necessary to sustain a glow discharge. It was hoped, in this manner, that some of the causes of variations of the electrical characteristics in a glow discharge would be revealed.

III. EXPERIMENTAL TECHNIQUE

Tubes were constructed as in Fig. 1; the test circuit is sketched next to the design of the tube.

All tubes were constructed with trap-door anodes. The purpose here was to provide uncontaminated anodes and to facilitate cathode temperature measurements. During exhaust, when the cathode was activated, the trap was in an upright position. After cathode activation, the trap door was dropped and processing continued.

It is to be noted that all temperatures were measured on the cathode base metal. Thus the cathode surface was estimated to be from 0 to 50 degrees cooler than the temperature of the base metal. For the experiments to be described, extreme exactness of temperature measurement at the cathode surface was not required.

The following procedure was established for testing. After aging, the sparking potential was measured and the temperature of the base metal recorded. By means of adjusting the power supply voltage the glow was brought so that it just covered the cathode surface, and did not flow down the sides of the cylinder. The area was taken as the area of the cathode, the current was measured and the current density calculated. The normal glow voltage was recorded by taking the fall in potential across the tube at this point. It was noted that the tube drop changed very little by varying the area of cathode covered by the glow discharge. After the normal glow tube drop was recorded, the current was decreased in the tube until the glow was extinguished. This was a rather abrupt point. Sometimes a

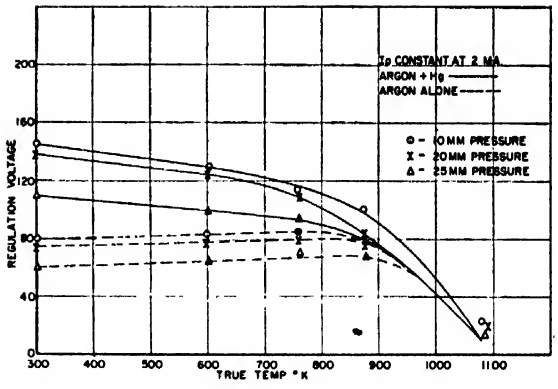


FIG. 3. Regulation voltages as a function of temperature for argon and argon contaminated with mercury.

sudden surge of current appeared as the glow was extinguished. The meaning here is not clear. Generally the current was decreased to the microampere region before it was extinguished. The minimum current refers to the lowest current that could be drawn before the glow became non-self-sustaining. The pressure of each tube was also recorded.

IV. DATA AND DISCUSSION CONCERNING SPARKING POTENTIALS AND REGULATION VOLTAGES

For each tube, a determination was made of sparking potentials, and regulation voltage, as a function of the temperature of the base metal. Some sample curves are to be found in Figs. 2 and 3. It is of interest to note that as the cathode temperature increases, the sparking potential decreases. This is to be expected to some extent. However, for argon, as the cathode temperature increases, from 300°K to about 750°K the regulation voltage does not fall, but, in fact, tends to show a slight rise. The general tendency or rising regulation voltage with rising cathode temperature was noted for all the tubes with the (BaSrCa)O cathodes and pure argon gas. Actually 19 tubes with this combination were tested and all showed the same tendency. Tubes made with neon gas and argon-neon mixture behaved similarly.

A question arises as to why the regulation voltage increased as the cathode temperature went from 300°K to about 750°K. By the accepted concepts, as the cathode becomes more efficient, liberating more electrons, the necessary ionization of the gas for a self-sustained discharge should be produced by a lower

TABLE I. Sparking potentials in argon.

Pressure times distance	Barium metal cathode (300°K)	(BaSrCa)O cathode (300°K)	(BaSrCa)O cathode (500°K)
1.0 mm×cm	92 volts	95 volts	90 volts
1.5	100	109	99
2.0	111	120	110
2.5	121	134	120
3.0	128	144	126

TABLE II. Minimum sustaining currents for oxide cathodes and argon gas.

Minimum sustaining current in ma at room temperature	Minimum sustaining current in ma at 500°K, cathode temperature	Pressure in mm of mercury
0.08	0.06	5
0.22	0.11	10
0.37	0.16	15
0.48	0.20	20
0.62	0.25	25
0.75	0.29	30

field. Yet, experimentally this was not the case. One might suspect further that electron liberation just prior to the spark has a different mechanism from electron liberation due to the glow, since in the glow there is an increase in the number of metastables and photons close to the cathodes.

Another experiment was then performed where tubes were made up with argon gas and (BaSrCa)O cathodes, but this time a droplet of mercury was added prior to exhaust. After exhaust a portion of the mercury was still present as a droplet. During testing it was apparent that the presence of mercury vapor at room temperature had changed the characteristics of the tubes markedly, in the color of the glow, in the electrical properties, and in the cathode activity. During testing it was noted that the Hg greatly deactivated the (BaSrCa)O. Sometimes, as a cold cathode, the spark and glow would occur from the nickel instead of the cathode. Sometimes, only a part of the cathode was covered by the glow. In the following data only those tubes were used in which the glow was on the oxide-coated cathode surface. It was found, for mercury contaminated tubes, that as cathode temperature was increased, the regulation voltage decreased. This effect was in contrast to pure argon or neon where with increasing temperature the regulation voltage increased. As a matter of side interest the regulation voltage was higher for the argon and Hg tubes. The actual data are shown in Fig. 3.

The experiment with Hg vapor indicated the marked effect of Hg vapor in changing the glow characteristics. But more than this, it offers a possible explanation for the rather anomalous behavior of the regulation voltage in pure argon. In the glow discharge the liberation of electrons is aided not only by positive ions striking the surface but largely by photons and metastables. The effects of metastable atoms is greater in the glow than in the conditions leading to the spark. As the cathode in a glow discharge is heated, one might expect the gas in the cathode vicinity to have an increased kinetic energy. It seems reasonable to expect that a metastable atom can lose its energy by collision with a neutral atom, with the excess energy given off by the kinetic energy of the particles involved. If the kinetic energy of the gas particle is increased, one might expect the metastables to be decreased, thus the process of liberation of electrons from the cathode would be less efficient and a higher field would be necessary to maintain the discharge. One might expect that for pure argon the regulation voltage would rise with cathode temperature until the temperature rose enough to emit sufficient electrons for the tube to go into an externally sustained arc. The presence of the mercury in argon would effectively reduce the metastable argon atoms by ionization of the mercury. Without the metastables being present, the regulation voltage should drop with increasing cathode temperature and this is what has been observed.

Another point of interest is the determination of sparking potentials and regulation voltages at various pressures. For the case of (BaSrCa)O cathodes, it was found that the curves for sparking potential *versus* pressure×distance were very similar to those determined with barium metal cathodes in the same gas[4] (Table I).

It has been found, experimentally, that the regulation voltage, at constant temperatures, changes very slightly with pressure. In some cases, there is a slight increase of regulation voltage at reduced pressures.

V. CURRENT DENSITY AND MINIMUM SUSTAINING CURRENT

The current density in a glow discharge has been studied theoretically and the following equation[5] is most often quoted to describe it:

$$_n = 5.92\times10^{-14}\frac{AB^2(k_+p)(1+\gamma)}{\log(1+1/\gamma)}p^2\frac{\text{amp}}{\text{cm}^2}. \qquad (1)$$

In the experiments performed, current density measurements were made with respect to pressure and cathode temperature. The current density was actually found to decrease with increasing temperature up to about 700°K. Above this temperature the current density would start to rise very rapidly. The data reveals

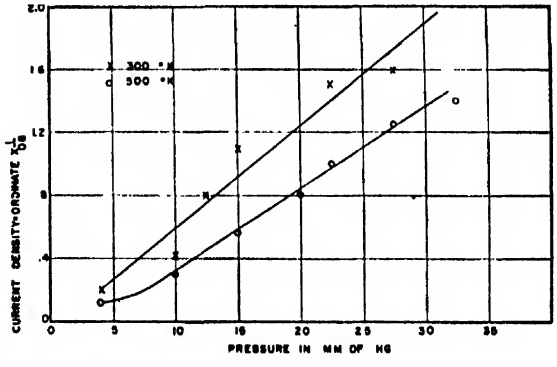


FIG. 4. Current density and pressure for argon-filled tubes and (BaSrCa)O cathodes. (Legend to the left of the figure should read: current density = ordinate × 1/0.08 ma/cm².)

[4] H. Jacobs and A. P. LaRocque, J. App. Phys. 18, 199–203 (1947).
[5] See reference 2, p. 584.

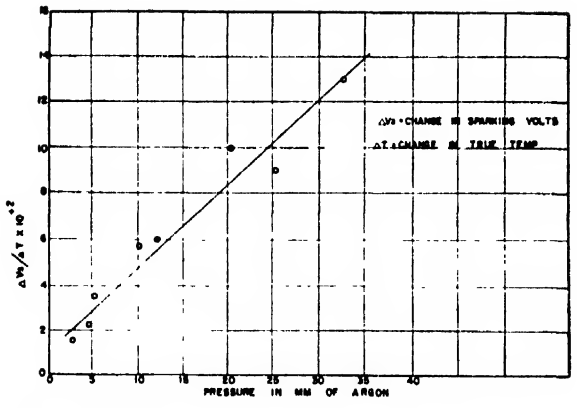


FIG. 5. The role of pressure on the change of sparking potential with respect to temperature (volts/°K) for argon gas and (BaSrCa)O cathodes.

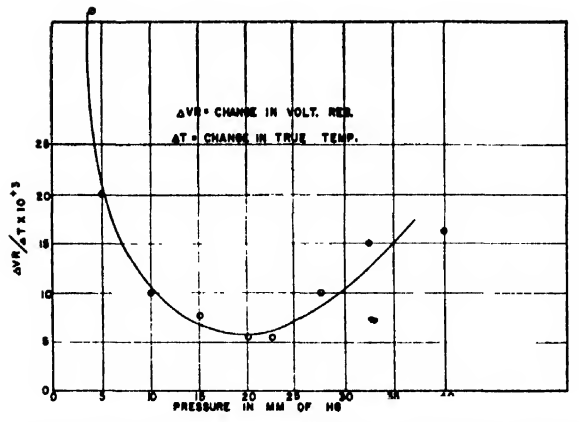


FIG. 6. The role of pressure on the change of regulation voltages with respect to temperature (volts/°K), for argon gas and (BaSrCa)O cathodes.

that as the cathode rises in temperature, the regulation voltage rises and the current density diminishes, until the region of the arc is approached, where the voltage drops rapidly and the current density rises rapidly.

Referring to Eq. (1) again, which was derived on space-charge considerations of positive ions, as p^2 increases, j_n should increase. In Fig. 4 we find the current density *vs.* pressure at 300°K temperature and at 500°K, for argon. This indicates two factors. (a) Current density decreases with increasing temperatures in this range. (b) Current density is almost a linear function of pressure. (For the mixtures of neon and argon, the current density was found to increase at a slightly higher rate as the pressure was increased.)

At the same period of measurement the minimum sustaining current was measured and found to decrease with increasing temperature (Table II).

VI. FURTHER CONSIDERATIONS

Upon study, the experiments with uncontaminated tubes indicated rather definite relationships between the change in sparking potential with respect to cathode temperature, dV_s/dT, and pressure. The change in regulation voltage with respect to pressure, dV_R/dT, was also found to be a function of pressure. As indicated in Fig. 5, dV_s/dT was found to be a constant as the pressure was increased.[6] Similarly (see Fig. 6) dV_R/dT was found to be a function of pressure—although not a constant.

The question then arises as to whether or not these experiments can give us a physical insight as to the mechanism in the spark. For instance, it has not been shown in what manner γ, the second Townsend coefficient, varies with cathode temperature for these particular cases. From Fig. 5, we can set up the following relation:

$$-dV_s/dT = Cp, \tag{2}$$

dV_s = change in sparking potential (volts), dT = change in cathode temperature (°K), C = constant, p = pressure in mm of mercury.

Using the Townsend criterion for sparking, Eq. (3) can be shown to be the equation for sparking.[7]

$$V_s = \frac{Bpd}{\ln[Apd/\ln(1/\gamma)]}, \tag{3}$$

where V_s = sparking potential, p = pressure, d = distance, γ = largely secondary emission on positive ion bombardment. This quantity, in addition, contains electrons liberated by photons and metastables. A and B = empirical constants.

Integrating Eq. (2),

$$-V_s = CpT + V_0. \tag{4}$$

V_0 is the extrapolated hypothetical sparking potential near 0°K for a given pressure and distance. We are now in a position to discuss the manner in which γ varies with cathode temperature. Substitute Eq. (3) in Eq. (4):

$$-\frac{Bpd}{\ln[Apd/\ln(1/\gamma)]} = CpT + V_0. \tag{5}$$

The following factors are assumed constant: B, p, d, A, C, and V_0. γ is the remaining factor varying with T.

$$T = -\frac{Bpd}{Cp\ln[Apd/\ln(1/\gamma)]} + \frac{V_0}{Cp}. \tag{6}$$

T varies as $K\{1/\ln[\ln(1/\gamma)]\}$, K = constant; $\ln[\ln(1/\gamma)]$ varies as K/T.

We can conclude: $\exp(e^{1/T})$ varies as $1/\gamma$ times a constant, for sparking.

Now let us refer to Fig. 6 which shows the increase of regulation voltage per change in cathode temperature as a function of pressure. We notice that at about 20

[6] $\Delta V_s/\Delta T$ is a negative quantity in Fig. 5.

[7] J. D. Cobine, *Gaseous Conductors* (McGraw-Hill Book Company, Inc., New York, 1941), p. 163.

mm, dV_R/dT is a minimum. The behavior here is not readily explained at the present time.[8]

VII. CONCLUSIONS

It was found that for experimental glow discharge tubes without mercury contaminations, and in the cathode temperature regions of from 300°K to 750°K, using oxide cathodes and inert gases:

1. The sparking potentials fall with increasing cathode temperatures.
2. The regulation voltage rises slightly with increasing cathode temperature. (For mercury vapor and argon, the regulation voltage falls with increasing cathode temperature.)
3. The current density decreases with increasing cathode temperature. This takes place concurrently with the increase in regulation voltage.
4. The minimum sustaining current in a glow discharge decreases with increasing cathode temperature, and rises with increasing pressure.
5. For the case of argon, the current density in the glow discharge varies roughly linearly with pressure in contrast with the space charge derivations which predict a variation with pressure squared. For the case of the neon, argon mixtures, the rate of increase in current density is not quite linear but slightly greater.
6. The decrease in sparking potential with respect to temperature diminished linearly with lower pressures. The change of regulation voltage with respect to temperature is related to pressure, but not in any simple manner.

[8] T. Jurrianse, Philips Res. Rep. 1, 407 (1946). For the case of neon gas and a molybdenum cathode, it was found that dV_R/dT was a negative quantity.

VIII. ACKNOWLEDGMENT

We wish to thank Mr. K. M. McLaughlin and the Radio Corporation of America Development Group at Harrison, New Jersey, for their suggestions in the tube design and for the construction of the experimental tubes.

The Emission of Radiation from Diatomic Gases. I. Approximate Calculations*

S. S. Penner
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California
(Received February 6, 1950)

An approximate method for estimating radiant heat transfer from gaseous emitters is proposed. An average absorption coefficient is used for the effective width of an entire vibration-rotation band. The procedure for determining an average absorption coefficient in terms of the integrated absorption is justified, approximately, for very small optical densities and also for large total pressures where the spectral half-width is no longer small compared with the rotational spacing. Because of these limitations it is to be expected that the procedure proposed here will be particularly useful only in estimating gaseous emissivities for emitters in high-pressure combustion chambers. A preliminary compilation is given of the theoretical results which are necessary for making rapid estimates of the emissivities, over a wide temperature range, for CO, HI, HBr, HCl, and HF.

I. INTRODUCTION

PROBLEMS concerned with the emission of radiant energy from an assembly of gaseous emitters which are in thermal equilibrium have formed the subject for theoretical and experimental investigations by two essentially distinct groups.

The members of one group, the theoretical physicists, have attacked the fundamental atomistic problem in an effort to evolve relations, based on quantum-mechanical considerations, which permit a rigorous solution of the problem, at least in principle. The basic relations between the various factors contributing to radiant heat transfer are correlated by the well-known relations between the coefficients of induced and spontaneous emission and induced absorption developed by Einstein in 1917.[1] Theoretical efforts to derive relations for the Einstein coefficients from first principles have culminated in Dirac's treatment of the interaction of radiation and matter, numerical application of which to specific problems usually involves formidable computational difficulties. Excellent reviews from the fundamental point of view of the work of Dirac, Born, Slater, and others have been given by Fermi[2] and Breit.[3] The subject of the interaction of radiation and matter is still far from solved, as is evidenced by intensive current activity in this field.[4]

Contributors of the second group involved in the calculation of intensities of radiation come from the ranks of the applied scientists and engineers who have contented themselves with a phenomenological description of the intensities of radiation emitted from gaseous bodies of various shapes, sizes, and compositions and who have based their predictions primarily upon empirically determined parameters such as effective over-all emissivities for pure gases and gaseous mixtures as functions of pressure, temperature, and path length. Although this approach has involved only

* This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. W-04-200-ORD-455, sponsored by the U. S. Army Ordnance Department.
[1] A. Einstein, Physik. Zeits. 18, 121 (1917).
[2] E. Fermi, Rev. Mod. Phys. 4, 87 (1932).
[3] G. Breit, Rev. Mod. Phys. 4, 504 (1932).
[4] See, for example, V. F. Weisskopf, Rev. Mod. Phys. 21, 305 (1949), and references cited in this article.

limited application of the fundamental relations determining the intensity of radiant energy emitted, it has nevertheless been developed to the point where reasonable predictions can be made for practical applications.[5]

It is apparent that use of the results of the theoretical physicists will suggest the most desirable lines of attack along which practical methods for the calculation of radiant heat transfer should be evolved. Accordingly, we shall devote our attention in this paper and in succeeding articles on radiant heat transfer to utilizing some of the established results relating to the calculation of radiant energy emitted from pure gases. For the present we shall content ourselves with developing and applying a simple and rather crude method for estimating gaseous emissivities by assigning an average absorption coefficient to an entire vibration-rotation band. Before proceeding with the presentation of details of this development, it appears desirable to review briefly the essential steps in estimating the intensity of radiation emitted from heated gases.

A gas emits radiation as the result of electronic, vibrational, and rotational transitions from excited energy levels to lower energy levels. The emitted radiant energy corresponding to these transitions is distributed over a wide wave-length region. Only the transitions corresponding to the infra-red vibration-rotation bands make an important contribution to the total radiant heat transfer at the temperatures of interest in connection with studies on combustion chambers. The total radiant intensity emitted from a given quantity of gas is obtained by summing the radiant intensities corresponding to each of the individual energy transitions. A calculation of this type requires knowledge of the following factors: 1. The number of molecules present in each of the various energy levels must be known. If thermal equilibrium obtains, this distribution can be calculated from quantum statistics. If thermal equilibrium does not exist, as may well be the case in a region of active combustion, then the equilibrium assumption and hence the numerical results derived from it are obviously invalid. 2. The transition probabilities for each of the possible energy transitions must be determined. Transition probabilities can be calculated with satisfactory precision for isolated quantum-mechanical systems such as vibrating-rotating diatomic molecules. At elevated temperatures and pressures, however, the various radiating systems may be perturbed as the result of interactions with neighboring molecules. These perturbations are neglected in the present calculations. 3. The frequencies corresponding to each of the energy transitions enter into the quantitative expressions for the radiant intensities. These frequencies can usually be determined with a high degree of precision from available spectroscopic data. 4. The spectral line shape must be known as a function of pressure and temperature. It can be calculated with relative ease if the familiar Lorentz collision formula is valid, and with some difficulty when it is not valid. In the present approximation, calculations of radiant heat transfer will be carried out without explicit allowance for line shape by introducing the restriction that the total pressure is sufficiently high to provide extensive overlapping of individual rotational lines.

II. BASIC RELATIONS BETWEEN INTEGRATED ABSORPTION, EINSTEIN COEFFICIENTS, AND RADIANT INTENSITIES

An energy transition from a lower energy level j to an upper energy level i (or conversely) is accompanied by the absorption (or emission) of radiation of frequencies close to the frequency ν_{ij} given by the familiar quantum relation $\nu_{ij}=\Delta E_{ij}/h$ where ΔE_{ij} is the energy difference between the two levels under consideration, and h is Planck's constant. Let $\rho(\nu_{ij})$ represent the volume density of radiant energy emitted by a blackbody in the frequency interval between ν_{ij} and $\nu_{ij}+d\nu_{ij}$. The integrated absorption $\int P_{\nu_{ij}}d\nu_{ij}$ is related to the Einstein coefficients of induced absorption and induced emission by the well-known relation[6]

$$\int P_{\nu_{ij}}d\nu_{ij}\simeq N_jB_{j\rightarrow i}h\nu_{ij}/c-N_iB_{i\rightarrow j}h\nu_{ij}/c = N_iA_{i\rightarrow j}h\nu_{ij}/c\rho(\nu_{ij}), \quad (1)$$

where N_j = number of molecules per unit volume in quantum state j, N_i = number of molecules per unit volume in quantum state i, $B_{j\rightarrow i}$ = Einstein coefficient of induced absorption, $B_{i\rightarrow j}$ = Einstein coefficient of induced emission, $A_{i\rightarrow j}$ = Einstein coefficient of spontaneous emission, and c = velocity of light. The symbol $P_{\nu_{ij}}$ indicates that the absorption coefficient with which we are concerned arises as the result of an energy transition between the levels i and j. The integration in Eq. (1) may be performed over all frequencies $-\infty<\nu_{ij}<+\infty$ if we adopt the convention of calling frequencies smaller than ν_{ij} negative and frequencies larger than ν_{ij} positive.

Treatments of intensities of emission and absorption involving the Einstein coefficients are derived under conditions of thermal equilibrium.[1,6,7] Nevertheless it is customary to assume[8] that energy transitions from the upper level i to the lower level j occur at the net rate,

$$N_i[A_{i\rightarrow j}+B_{i\rightarrow j}\rho'(\nu_{ij})]-N_jB_{j\rightarrow i}\rho'(\nu_{ij}), \quad (2)$$

even if thermal equilibrium does not exist. Here $\rho'(\nu_{ij})$

[5] H. C. Hottel, "Radiant heat transfer," in W. H. McAdam's *Heat Transmission* (McGraw-Hill Book Company, Inc., New York, 1942); M. Jakob, *Heat Transfer* (John Wiley and Sons, Inc., New York, 1949).

[6] R. C. Tolman, Phys. Rev. 23, 693 (1924); E. G. Kemble, Phys. Rev. 25, 1 (1925).

[7] L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1935), p. 299.

[8] E. C. Kemble, *Fundamental Principles of Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1937), p. 450.

is the external energy density of radiation passing through the system under discussion.

An expression of the type given in Eq. (2) is not particularly useful for calculations of the net radiant intensity emitted. Instead we require a differential expression which can be obtained by considering the change of energy resulting from the passage of radiant energy through a radiating and absorbing system of emitters and receivers. If the incident radiant intensity (not necessarily the intensity of radiation emitted by a blackbody) is denoted as $I(\nu)$, the following differential equation results

$$dI(\nu)=P_\nu[c\rho(\nu)-I(\nu)]dX.$$

Here it is assumed that the fraction of radiant energy absorbed is proportional to the incident intensity of radiation (with the proportionality constant set equal to the spectral absorption coefficient P_ν) and that the net change in radiant energy is proportional to the change in optical density dX through the system of emitters and absorbers. Integration leads directly to the well-known relation

$$I(\nu)=c\rho(\nu)[1-\exp(-P_\nu X)] \qquad (3)$$

for the radiant energy emitted per unit area of surface by gaseous emitters distributed uniformly through a radiation path length of optical density X.

Equation (3) cannot be applied to practical calculations unless the dependence of P_ν on ν is known, a problem which requires detailed consideration of spectral line shape and its dependence on pressure, temperature, etc.[9,10] If Lorentz broadening obtains, P_ν is related to the integrated absorption by the dispersion formula

$$P_\nu=\frac{\delta}{\pi}\frac{\int P_{\nu,ij}d\nu_{ij}}{(\nu-\nu_{ij})^2+\delta^2}, \qquad (4)$$

where δ represents the spectral half-width. If individual rotational lines are sufficiently broadened to cause overlapping of lines, then the spectral absorption coefficient, for collision-broadened lines, is given by a sum of terms of the type shown in Eq. (4). Therefore, in this particular case, the intensity of radiation $I(\nu)$ can be calculated provided the half-widths δ and the integrated absorption $\int P_{\nu,ij}d\nu_{ij}$ are known for each transition. The half-width δ can be determined, for example, from the dependence of absorption on pressure[11] by utilizing the theoretical analysis of Elsasser.[12,13]

TABLE I. Effective band width of the fundamental of CO as a function of temperature.

T °K	j_{max}	ν_{max} cm^{-1}	ν_{min} cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
300	28	2235	2021	214
500	36	2256	1982	274
1000	50	2287	1908	379
1500	62	2306	1841	465
2000	72	2317	1782	535
2500	80	2322	1732	590
3000	86	2325	1694	630

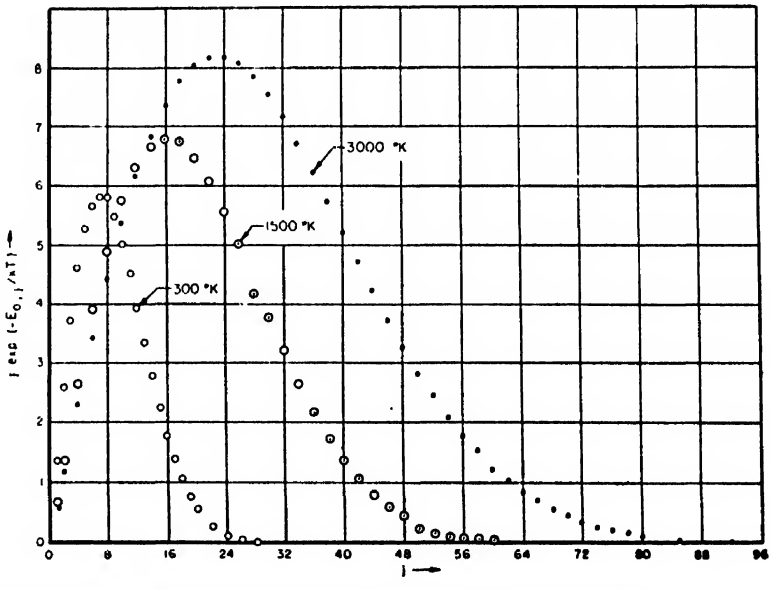


FIG. 1. Representative plots of $j\exp(-E_{0,j}/kT)$ *vs.* j for carbon monoxide (different scales are used for the ordinate at different temperatures).

[9] H. Margenau and W. W. Watson, Rev. Mod. Phys. 8, 22 (1936); H. M. Foley, Phys. Rev. 69, 616 (1946).
[10] J. H. Van Vleck and V. F. Weisskopf, Rev. Mod. Phys. 17, 227 (1945); P. W. Anderson, Phys. Rev. 76, 647 (1949); E. Lindholm, Dissertation, Uppsala (1942).
[11] M. Summerfield and J. Strong, Phys. Rev. 59, 217 (1941); A. M. Thorndike, J. Chem. Phys. 16, 211 (1948).
[12] W. M. Elsasser, Phys. Rev. 54, 126 (1938); Astrophys. J. 87, 497 (1938).
[13] W. M. Elsasser, Harvard Meteorological Studies No. 6, 1942. The theoretical treatment of Elsasser has recently been generalized to arbitrarily spaced rotational lines of arbitrary intensity distribution [see F. Matossi, R. Mayer, and E. Rauscher, Phys. Rev. 76, 760 (1949)].

The dependence of δ on pressure and temperature can be calculated, approximately, for collision-broadened lines.[9,10] Individual values for integrated absorption for rotational transitions can be calculated if the integrated absorption for a vibration-rotation band has been determined experimentally. Calculations for diatomic molecules involve the use of theoretical results obtained by Oppenheimer[14] and applied by Dennison.[15] Similar calculations are also possible for the simpler polyatomic molecules.

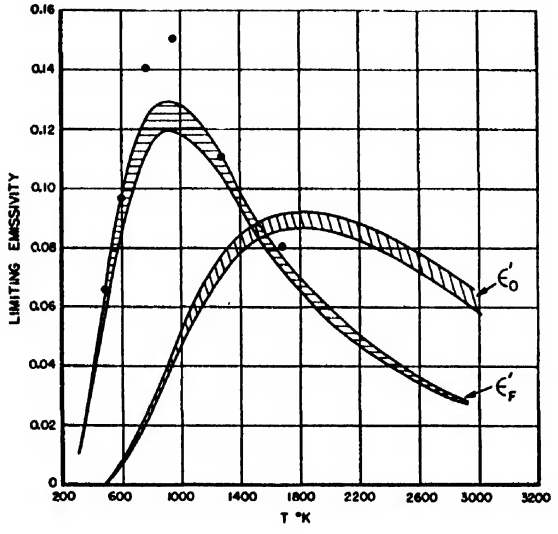


FIG. 2. Limiting emissivity of CO as a function of temperature Experimental data for $pl = 5$ ft-atmos. (cf. reference 19): ●

T °K	j_{max} (upper curve)	j_{max} (lower curve)
300	30	28
500	38	36
1000	54	50
1500	66	62
2000	76	72
2500	86	80
3000	92	86

III. A METHOD FOR THE APPROXIMATE CALCULATION OF THE INTENSITY OF RADIATION EMITTED FROM GASES WHICH ARE IN THERMAL EQUILIBRIUM

In this section we shall consider drastic simplifications of methods for making radiant heat-transfer calculations for gases.

A. A Method for Estimating an Upper Limit for the Emission of Radiant Energy from Pure Gases and Gaseous Mixtures

An upper limit is obtained for the total radiant energy $\int_{n\to n'} I_\nu d\nu$ emitted for frequencies corresponding to the vibrational transition from n to n' if we set

$$\int_{n\to n'} I_\nu d\nu = c\int_{\Delta\nu_{n\to n'}} \rho(\nu)d\nu, \quad . \quad . \quad (5)$$

where $\Delta\nu_{n\to n'}$ designates an appropriate band width for the vibration-rotation band under consideration. The total radiant energy emitted over all emission bands is evidently obtained by summing the contributions for all possible vibrational transitions, i.e., The maximum

$$I_T = \sum_{n,n'}\int_{n\to n'} I_\nu d\nu = c\sum_{n,n'}\int_{\Delta\nu_{n\to n'}} \rho(\nu)d\nu. \quad (6)$$

value of the effective emissivity ϵ'_{max} is defined by the relation

$$\epsilon'_{max} = \left\{\sum_{n,n'}\int_{\Delta\nu_{n\to n'}} \rho(\nu)d\nu\right\} \Big/ \int_0^\infty \rho(\nu)d\nu. \quad (7)$$

The calculation of an upper limit for the emissivity is seen to require an independent estimate of band widths $\Delta\nu_{n\to n'}$. It is to be expected that a suitable method for the calculation of $\Delta\nu_{n\to n'}$ will lead to theoretical upper limits for the emissivities which are in agreement with the experimentally observed limits approached by given emitters as the optical density is increased. The method suggested here for the calculation of effective band width approximately fulfills this requirement.[16] The principal factor determining the dependence of integrated absorption on rotational quantum number is of the form

$$j\exp(-E_{n,j}/kT) \quad \text{or} \quad j\exp(-E_{n,j-1}/kT),$$

where $E_{n,j}$ represents the energy of the nth vibrational and jth rotational level, k is the Boltzmann constant, and T represents the absolute temperature. These facts suggest that the band width $\Delta\nu_{n\to n'}$ be set equal to the frequency interval for which the factors $j\exp(-E_{n,j}/kT)$ or $j\exp(-E_{n,j-1}/kT)$ have values in excess of 10^{-3} of their maxima at a given temperature. This definition of $\Delta\nu_{n\to n'}$ is admittedly arbitrary. However, the exponential terms decrease so rapidly with increasing values of j that a different definition, such as setting $\Delta\nu_{n\to n'}$ equal to the frequency interval in which $j\exp(-E_{n,j}/kT)$ or $j\exp(-E_{n,j-1}/kT)$ have values in excess of 10^{-2} or 10^{-4} of their respective max-

TABLE II. Effective band width of the first overtone of CO as a function of temperature.

T °K	j_{max}	ν'_{max} cm^{-1}	ν'_{min} cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
300	28	4337	4124	213
500	36	4349	4078	271
1000	50	4357	3981	376
1500	62	4357	3890	467
2000	72	4357	3808	549
2500	80	4357	3737	620
3000	86	4357	3682	675

[14] J. R. Oppenheimer, Proc. Camb. Phil. Soc. 23, 327 (1926).
[15] D. M. Dennison, Phys. Rev. 31, 503 (1928).

[16] The utility of an empirically determined effective band width seems to be supported by absorption measurements performed by the author and D. Weber. The "effective band width" of the CO fundamental at room temperature was found to increase only about 20 percent when the pressure of absorbing gas was increased from 5 mm to 200 p.s.i. at constant total pressure.

ima, would lead to substantially the same results as the definition of band width adopted for the present application. It should be noted that the definition of effective band width given here leads to similar results for all transitions of the form $n \rightarrow n+1$ provided the anharmonicities are not excessive. Estimates of overtone intensities indicate that the effective band widths of the overtones correspond, approximately, to the same limiting values for the rotational quantum numbers j as tor the fundamental

The foregoing considerations lead to a straightforward method for estimating effective band widths. The application of this method is described here for carbon monoxide.

In Fig. 1 a plot is shown of $j\exp(-E_{0,j}/kT)$ for carbon monoxide as a function of j for representative values of the temperature. Similar plots have been prepared for each of the following temperatures: 300, 500, 1000, 1500, 2000, 2500, and 3000°K. Reference to Fig. 1 shows that for all values of j up to about $j_{\max}=28$ the integrated absorption is larger than 10^{-3} of its maximum value at 300°K. Similar results are obtained at other temperatures. It follows from the preceding discussion that $\Delta\nu_{0\rightarrow1}\simeq\Delta\nu_{1\rightarrow2}\simeq\Delta\nu_{2\rightarrow3}\simeq\cdots=$ effective band width of the fundamental $=\nu_{\max}-\nu_{\min}$, where $\nu_{\max}$ and $\nu_{\min}$ are given, respectively, by the approximate relations

$$\nu_{\max}\simeq\frac{E(1,j_{\max})-E(0,j_{\max}-1)}{h} \tag{8}$$

$$\nu_{\min}\simeq\frac{E(1,j_{\max}-1)-E(0,j_{\max})}{h}. \tag{8a}$$

Band widths have been calculated for the fundamental vibration-rotation band of carbon monoxide from available spectroscopic data[17] by the use of this method. The results are summarized in Table I.[18]

For sufficiently large values of j a band head should be obtained beyond which the calculated maximum cutoff frequency should decrease with increasing values of $j_{\max}$. This phenomenon is observed for smaller j-values in the case of the first overtone of CO (see below). When a band head does occur, the cut-off frequency should evidently be chosen to coincide with the extreme frequency value corresponding to the band head rather than to a frequency calculated from either Eq. (8) or Eq. (8a).

In accord with the preceding discussion we set $\Delta\nu_{0\rightarrow2}\simeq\Delta\nu_{1\rightarrow3}\simeq\cdots=$ effective band width of the first overtone $=\nu'_{\max}-\nu'_{\min}$ where

$$\nu'_{\max}\simeq\frac{E(2,j_{\max})-E(0,j_{\max}-1)}{h} \tag{9}$$

$$\nu'_{\min}\simeq\frac{E(2,j_{\max}-1)-E(0,j_{\max})}{h}. \tag{9a}$$

As explained in the preceding paragraph, Eq. (9) or Eq. (9a) should not be used if a band head is observed for j values which are smaller than the appropriate values of $j_{\max}$. Relevant band width calculations for the first overtone of CO are summarized in Table II.

Since the integrated absorption of the second overtone of CO is very small compared with that of the fundamental and since, furthermore, the intensity of radiation emitted by a blackbody at temperatures below about 2000°K is relatively small in the frequency region in which the second overtone emission band is found, the contributions of the second overtone to radiant heat transfer will not become important until relatively high temperatures and large optical densities have been reached. These contributions may be included in radiant heat-transfer calculations by following the procedure outlined above for calculations on the fundamental and the first overtone.

If only the fundamental and the first overtone make significant contributions to the intensity of radiation emitted from a given gas, then Eq. (7) reduces to the relation

$$\epsilon'_{\max}=\frac{\sum_n\int_{\Delta\nu_{n\rightarrow n+1}}\rho(\nu)d\nu}{\int_0^\infty\rho(\nu)d\nu}+\frac{\sum_n\int_{\Delta\nu_{n\rightarrow n+2}}\rho(\nu)d\nu}{\int_0^\infty\rho(\nu)d\nu}=\epsilon_F'+\epsilon_0'. \tag{10}$$

The ratios ϵ_F' and ϵ_0' occurring on the right-hand side of Eq. (10) have been evaluated by using the effective band widths summarized in Tables I and II. These ratios have also been calculated for effective band widths determined by making a somewhat different choice for $j_{\max}$ at each temperature. The results of these calculations are plotted in Fig. 2. Reference to Fig. 2 indicates that the contribution to the radiant intensity emitted from the fundamental ϵ_F' shows a maximum value near 900°K. The fraction of the total radiant energy contributed by the first overtone ϵ_0' reaches a maximum value at temperatures in the neighborhood of 2000°K. At still higher temperatures contributions from upper harmonics will become important.

[17] H. Sponer, Molekülspektren (Verlag Julius Springer, Berlin, 1935); J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1940), pp. 468–469; for more recent spectroscopic data on CO, see G. Herzberg and K. N. Rao, J. Chem. Phys. 17, 1099 (1949).

[18] Band widths and corresponding emissivities have also been calculated for values of $j_{\max}$ somewhat different from those listed in Table I. The effect of the choice of $j_{\max}$, within reasonable limits, on the calculated emissivity was found to be small and is illustrated in Fig. 2.

According to the present procedure, the upper limit for the theoretical emissivity ϵ'_{max} is obtained by adding the contributions from the fundamental ϵ_F' and the first overtone ϵ_0' according to Eq. (10). It should be noted, however, that there is a certain inconsistency inherent in the use of a curve corresponding to the sum of the emissivity contributions from the fundamental and the first harmonic. Thus the integrated overtone absorption coefficient for a given rotational line is very small ($\sim 10^{-2}$ for CO) compared with the corresponding absorption coefficient of a rotational line of the fundamental. Therefore, the overtone will not act as a blackbody absorber or emitter until the optical density is exceedingly large. But for optical densities sufficiently large to make the overtone act as a blackbody, the intensity of radiation from the fundamental, which makes an appreciable contribution to the total radiant heat transfer, may be appreciable even for frequencies outside the range of the calculated band widths. Hence when the overtone acts as a blackbody, the fundamental must have an emissivity greater than the emissivity shown in Fig. 2.

Remarks similar to those made for radiant heat transfer from the first overtone apply to radiant heat transfer from upper harmonics.

TABLE III. Effective band width of the fundamental of HI as a function of temperature.

T °K	j_{max}	ν_{max} cm^{-1}	ν_{min} cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
300	16	2387	1988	399
500	22	2420	1880	540
1000	31	2439	1703	736
1500	37	2439	1577	862
2000	46	2439	1379	1060
2500	49	2439	1311	1128
3000	54	2439	1196	1243

TABLE IV. Effective band width of the first overtone of HI as a function of temperature.

T °K	j_{max}	ν'_{max} cm^{-1}	ν'_{min} cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
300	16	4513	4119	394
500	22	4514	3974	540
1000	31	4514	3720	794
1500	37	4514	3528	986
2000	46	4514	3208	1306
2500	49	4514	3093	1421
3000	54	4514	2894	1620

TABLE V. Effective band width of the fundamental of HBr as a function of temperature.

T °K	j_{max}	ν_{max} cm^{-1}	ν_{min} cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
300	15	2755	2261	494
500	19	2785	2167	618
1000	27	2813	1964	849
1500	33	2814	1801	1013
2000	41	2814	1575	1239
2500	43	2814	1517	1297
3000	47	2814	1400	1414

The preceding considerations lead to the conclusion that for moderate values of the optical density it is to be expected that observed emissivity values will tend to approach and then slightly exceed the emissivity corresponding to the emission of radiant energy from the fundamental, with the fundamental acting as a blackbody emitter. This qualitative conclusion is borne out by experimental data of the emissivity of carbon monoxide measured by Hottel and Ullrich,[5,19] which are also shown in Fig. 2. The experimental data for an optical density corresponding to 5 ft.×atmos. are seen to follow the calculated blackbody emissivity curve of the fundamental with reasonable accuracy. In view of the approximations involved in the present calculations, the agreement between theory and experiment must be considered to be satisfactory.

In concluding this discussion it appears desirable to point out that the method for estimating upper limits for the emissivities described here lends itself readily to calculations on polyatomic molecules.

B. A Method for Estimating Emissivities of Pure Gases

A more realistic estimate of the emissivity of pure gases than is provided by Eq. (10) is obtained by using an appropriate average absorption coefficient for each of the vibration-rotation bands. In this case Eq. (10) should be replaced by the following relation for the emissivity ϵ':

$$\epsilon' = [1-\exp(-\bar{k}_F pl)]\frac{\sum_n \int_{\Delta\nu_{n\to n+1}} \rho(\nu)d\nu}{\int_0^\infty \rho(\nu)d\nu} + [1-\exp(-\bar{k}_{F.O.}pl)]\frac{\sum_n \int_{\Delta\nu_{n\to n+2}} \rho(\nu)d\nu}{\int_0^\infty \rho(\nu)d\nu} + [1-\exp(-\bar{k}_{S.O.}pl)]\frac{\sum_n \int_{\Delta\nu_{n\to n+3}} \rho(\nu)d\nu}{\int_0^\infty \rho(\nu)d\nu} + \cdots, \quad (11)$$

where $\bar{k}_F$ = average absorption coefficient for fundamental vibration-rotation band, $\bar{k}_{F.O.}$ = average absorption coefficient for first overtone, $\bar{k}_{S.O.}$ = average absorption coefficient for second overtone, p = partial pressure of absorber, and l = optical path length. If only the contributions to the total radiant heat transfer from the fundamental and from the first overtone are

[19] W. Ullrich, Thesis, Massachusetts Institute of Technology, Cambridge (1935).

of importance, it is evidently sufficient to use the first two terms on the right-hand side of Eq. (11). Numerical evaluation of ϵ' is facilitated by the use of the black-body intensity ratios over the effective band widths, i.e., ϵ_P' and ϵ_0', of the type presented in Fig. 2. The problem at hand therefore reduces to the evaluation of average absorption coefficients.

Before attempting to determine the numerical value of an average absorption coefficient it is appropriate to inquire about the existence of a coefficient of this type for an entire vibration-rotation band. A detailed consideration of this problem has indicated that for moderate values of the optical density pl an average absorption coefficient of the order of one-half the maximum absorption coefficient obtained with low spectral resolution should be observed provided P_ν can be represented as a linear function of ν.[20] The physical significance of a relation of the form, $\bar{k}=\frac{1}{2}P'_{\max}$ where $P'_{\max}$ is the maximum absorption coefficient observed under low resolution, is dubious. Of more obvious significance would be a relation between $\bar{k}$ and the integrated absorption of a given vibration-rotation band. A relation of this type can be derived by proceeding in the manner described next.

From Eq. (3) we obtain, for example, for the entire fundamental vibration-rotation band of a diatomic molecule with P and R branches but no Q branch, the relation

$$\sum_n \sum_j \int_{\substack{n\to n+1\\ \Delta j=\pm 1}} I_\nu d\nu = c\sum_n \sum_j \int_{\substack{\Delta\nu_{n\to n+1}\\ \Delta j=\pm 1}} \rho(\nu)[1-\exp(-P_\nu pl)]d\nu \quad (12)$$

where P_ν is given by Eq. (4) for collision-broadened lines. The subscript $n\to n+1$ indicates the type of vibrational transition, whereas the subscript $\Delta j=\pm 1$ identifies the rotational transitions. If the vibration-rotation band is sufficiently narrow, we may replace $\rho(\nu)$ by an average value over the effective band width $\Delta\nu_{n\to n+1}$. Denoting the product of (a) this constant value of $\rho(\nu)$, (b) the effective band width, and (c) the velocity of light by I_P^0, and denoting the intensity sum over the band branches on the left-hand side of Eq. (12) by I_P, we obtain the relation

$$\frac{I_P}{I_P^0}=\frac{1}{\Delta\nu_{n\to n+1}}\sum_n \sum_j \int_{\Delta\nu_{n\to n+1}} [1-\exp(-P_\nu pl)]d\nu. \quad (13)$$

For sufficiently small values of the optical density $P_\nu pl \ll 1$ for all values of ν. In this case we may expand the exponential in Eq. (13), retaining only the first two terms, and obtain the result

$$\frac{I_P}{I_P^0}=\frac{pl}{\Delta\nu_{n\to n+1}}\sum_n \sum_j \left(\int_{\substack{n\to n+1\\ j\to j-1}} P_\nu d\nu + \int_{\substack{n\to n+1\\ j-1\to j}} P_\nu d\nu\right) \quad (14)$$

[20] Compare A. Schack, Zeits. der techn. Physik 5, 267 (1924).

TABLE VI. Effective band width of the first overtone of HBr as a function of temperature.

T °K	$j_{\max}$	$\nu'_{\max}$ cm^{-1}	$\nu'_{\min}$ cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
300	15	5172	4685	487
500	19	5173	4559	614
1000	27	5173	4270	903
1500	33	5173	4024	1149
2000	41	5173	3658	1515
2500	43	5173	3561	1612
3000	47	5173	3359	1814

TABLE VII. Effective band width of the fundamental of HCl as a function of temperature.

T °K	$j_{\max}$	$\nu_{\max}$ cm^{-1}	$\nu_{\min}$ cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
300	13	3088	2553	535
500	17	3129	2438	691
1000	24	3166	2221	945
1500	31	3168	1990	1178
2000	35	3168	1854	1314
2500	37	3168	1784	1384
3000	45	3168	1503	1665

TABLE VIII. Effective band width of the first overtone of HCl as a function of temperature.

T °K	$j_{\max}$	$\nu'_{\max}$ cm^{-1}	$\nu'_{\min}$ cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
300	13	5831	5304	527
500	17	5835	5152	683
1000	24	5835	4851	984
1500	31	5835	4500	1335
2000	35	5835	4282	1553
2500	37	5835	4168	1667
3000	45	5835	3685	2150

The double sum in Eq. (14) evidently represents the integrated absorption over the fundamental vibration-rotation band. Relations similar to Eq. (14) can be derived for harmonic vibration-rotation bands.

By comparing Eq. (11) with Eq. (14) we obtain the following relations for the average absorption coefficients for sufficiently small values of the optical density:

$$\bar{k}_P=\frac{1}{\Delta\nu_{n\to n+1}}\sum_n \sum_j \left(\int_{\substack{n\to n+1\\ j\to j-1}} P_\nu d\nu + \int_{\substack{n\to n+1\\ j-1\to j}} P_\nu d\nu\right)=\frac{\alpha_{01}}{\Delta\nu_{n\to n+1}} \quad (15)$$

$$\bar{k}_{P.O.}=\frac{1}{\Delta\nu_{n\to n+2}}\sum_n \sum_j \left(\int_{\substack{n\to n+2\\ j\to j-1}} P_\nu d\nu + \int_{\substack{n\to n+2\\ j-1\to j}} P_\nu d\nu\right)=\frac{\alpha_{02}}{\Delta\nu_{n\to n+2}} \quad (16)$$

etc., where α_{01} and α_{02} represent the integrated absorp-

TABLE IX. Effective band width of the fundamental of HF as a function of temperature.

T °K	j_{max}	ν_{max} cm^{-1}	ν_{min} cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
500	12	4305	3343	962
1000	16	4359	3110	1249
1500	20	4374	2866	1508
2000	23	4374	2678	1696
2500	27	4374	2423	1951
3000	29	4374	2295	2079
3500	31	4374	2167	2207
4000	33	4374	2040	2334
4500	36	4374	1851	2523
5000	39	4374	1666	2708

TABLE X. Effective band width of the first overtone of HF as a function of temperature.

T °K	j_{max}	ν'_{max} cm^{-1}	ν'_{min} cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
500	12	7938	6995	943
1000	16	7938	6676	1262
1500	20	7938	6320	1618
2000	23	7938	6032	1906
2500	27	7938	5621	2317
3000	29	7938	5405	2533
3500	31	7938	5183	2755
4000	33	7938	4955	2983
4500	36	7938	4603	3335
5000	39	7938	4241	3697

tion over the fundamental and the first overtone, respectively.

The validity of Eqs. (15) and (16) has been demonstrated only under the very restrictive condition that $P_\nu pl \ll 1$ for all ν. This result will not be of any particular practical significance. Fortunately, however, it can be shown that Eqs. (15) and (16) also represent reasonable approximations when the total pressure is sufficiently high. This result can be obtained by utilizing Elsasser's theoretical treatment for equally intense and equally spaced rotational lines.[12,13] Elsasser has shown that for this idealized rotational distribution the fraction of radiation absorbed as the result of a given transition is

$$\frac{I}{I_0}=1-\exp\left[\frac{1}{d}\int P_{\nu_{ij}}d\nu_{ij}\tanh\left(\frac{2\pi\delta}{d}\right)pl\right] \qquad (17)$$

where d represents the spacing between rotational lines. For equally intense and equally spaced rotational lines, to which Elsasser's treatment applies, the following identity must be valid

$$(1/d)\int P_{\nu_{ij}}d\nu_{ij}=\alpha_{01}/\Delta\nu_{n\to n+1}. \qquad (18)$$

From Eqs. (17) and (18) it may evidently be inferred that relations of the form of Eqs. (15) and (16) apply for the average absorption coefficients provided δ is sufficiently large compared to d, a result which will obtain for large total pressures. In fact, as was first pointed out by Elsasser,[13] an absorption coefficient equal to the ratio of integrated absorption to line spacing represents a good approximation provided the ratio of half-width to line spacing is greater than or equal to $1/\pi$. Equations (15) and (16) exemplify this result as applied to an entire vibration-rotation band. We therefore arrive at the important conclusion that average absorption coefficients, as given in Eqs. (15) and (16) represent a crude but approximately justifiable result for sufficiently large total pressures. The condition $\delta/d \geq 1/\pi$ will be fulfilled for many molecules at pressures in excess of a few hundred pounds per square inch. Thus the present results are seen to be immediately applicable to liquid-fuel and solid-fuel rocket motors operating at pressures of a few hundred to several thousand pounds per square inch provided line broadening is still accounted for by the dispersion formula. For the present purposes the assumption of collision broadening and, therefore, of the approximate validity of Eqs. (15) and (16) for approximate heat-transfer calculations at elevated pressures will be made.[21]

The principal practical value of relations of the form of Eqs. (15) and (16) is the result of the fact that the integrated absorption is approximately independent of temperature and pressure for a given number of emitters or absorbers. Therefore the temperature dependence of the average absorption coefficient can be calculated as soon as the effective band width is known. In this connection it may be of interest to note that the present method of calculation is, to some extent, self-compensating as far as errors introduced by a poor estimate of band width are concerned. Thus a small value of $\Delta\nu$ will be partly compensated by a large value of $\bar{k}$.

The total integrated absorption (neglecting anharmonicities) corresponding to a transition from the $(n-1)$th to the nth vibrational energy level is determined by the following approximate relation[15]

$$\alpha_{01}=\frac{\pi\epsilon_P^2 N_T n\exp(-E_{n-1}/kT)}{3\mu c\sum_n^\infty \exp(-E_{n-1}/kT)}[1-\exp(-h\nu_0/kT)]$$

where ϵ_P represents the effective charge, N_T is the total number of absorbing (or emitting) molecules per unit volume per unit pressure, μ is the reduced mass, and ν_0 is the frequency of the band center. Since, approximately

$$\frac{\exp(-E_{n-1}/kT)}{\sum_{n=1}^\infty \exp(-E_{n-1}/kT)}=\frac{\exp[-(n-1)h\nu_0/kT]}{\sum_{n=1}^\infty \exp[-(n-1)h\nu_0/kT]}$$

[21] Unpublished absorption measurements on the fundamental of CO at pressures up to 800 p.s.i. and at room temperature have led to experimental results which are described by Eq. (4). Details concerning this work will be published elsewhere.

it follows that

$$\alpha_{01}=\pi\epsilon_F^2 N_T/3\mu c. \tag{19}$$

Relations similar to Eq. (19) can be derived readily for diatomic molecules with Q branch and also for polyatomic molecules.[22] Equation (15) may be written, in terms of the effective charge, as follows:

$$\bar{k}_P=\frac{N_T\pi\epsilon_F^2}{3\mu c}\frac{1}{\Delta\nu_{n\to n+1}}. \tag{20}$$

Since N_T varies inversely as the temperature, i.e., for an ideal gas

$$N_T=\frac{7.34\times10^{21}}{T}\text{ molecules/cc}\times\text{atmos.} \tag{21}$$

it follows that

$$\frac{\bar{k}_P(T_1)}{\bar{k}_P(T_2)}=\frac{T_2\,\Delta\nu_{n\to n+1}(T_2)}{T_1\,\Delta\nu_{n\to n+1}(T_1)}, \tag{22}$$

$$\frac{\bar{k}_{P.O.}(T_1)}{\bar{k}_{P.O.}(T_2)}=\frac{T_2\,\Delta\nu_{n\to n+2}(T_2)}{T_1\,\Delta\nu_{n\to n+2}(T_1)}, \tag{23}$$

etc. Thus the average absorption coefficient at any desired temperature can be estimated from available room temperature measurements by the use of Eqs. (22) and (23).

C. Critique of the Approximate Method for Emissivity Calculations

The principal virtue of the suggested procedure for emissivity calculations is the result of the fact that semiquantitative information regarding gaseous emissivities at elevated total pressures can be obtained by very simple calculations without requiring more than highly approximate values for the integrated absorption. Thus the integrated absorption for the fundamental vibration-rotation bands is generally so high that for total pressures of several hundred p.s.i. or more the fundamental will act practically as a blackbody at relatively low partial pressures of absorber. At elevated temperatures somewhat larger partial pressures of absorber will be required because of an increase in the effective band width with temperature. As the result of rapid saturation for absorption and emission from the fundamental, calculated emissivities tend to be relatively insensitive to the numerical values of the average absorption coeficients, at least as far as emission from the fundamental is concerned. The contributions from the first overtone to radiant heat transfer do not become important until very high partial pressures of absorber and relatively high temperatures have been reached.

Although emissivity calculations are relatively insensitive to the numerical values of the average absorption coefficient, they depend strongly on the numerical value of the effective band width. It is indeed

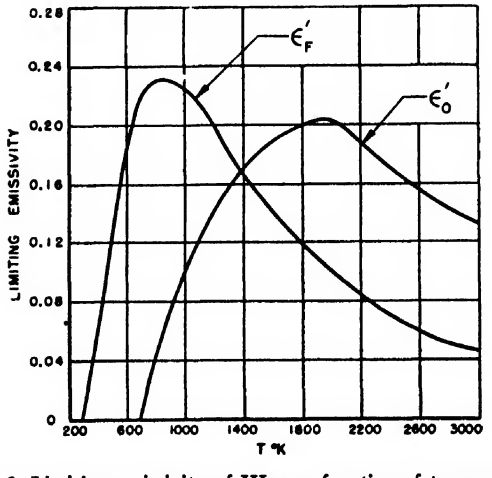


FIG. 3. Limiting emissivity of HI as a function of temperature.

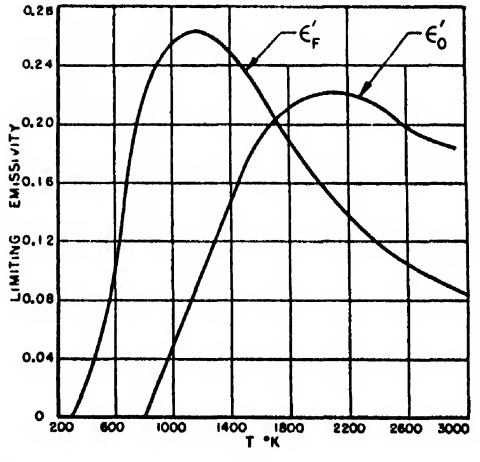


FIG. 4. Limiting emissivity of HBr as a function of temperature.

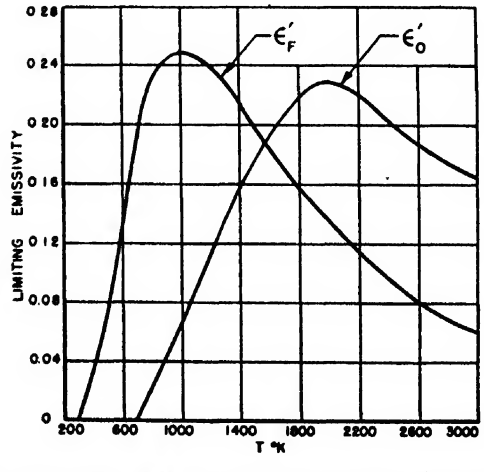


FIG. 5. Limiting emissivity of HCl as a function of temperature.

[22] M. W. Wolkenstein, J. Phys. USSR 5, 185 (1941); A. M. Thorndike, A. J. Wells, and E. B. Wilson, Jr., J. Chem. Phys. 15, 157 (1947); A. M. Thorndike, J. Chem. Phys. 15, 868 (1947).

fortunate that the effective band width does not vary rapidly with the pressure of absorbing gas.[16] This conclusion follows most obviously from an examination of the term $\exp(-P_e pl)$. The exponent is seen to be proportional to the optical density (pl) but changes with rotational quantum number as $j\exp(-E_{o,j}/kT)$. Since the effective band width is relatively insensitive to optical density, we are justified in concluding that curves of the limiting emissivity as a function of temperature for the fundamental vibration-rotation band, of the type shown in Fig. 2 for CO, represent a valuable indication of the semiquantitative variation of emissivity with temperature over a wide range of partial pressures of absorber at elevated total pressure and constant path length. As the partial pressure of the absorber is increased, the expected emissivity will exceed the limiting emissivity of the fundamental by roughly the amount contributed by the second term on the right-hand side of Eq. (11).

IV. APPLICATION TO DIATOMIC MOLECULES

The spectral emissivity for a number of diatomic gases can be determined, approximately, from available experimental data on integrated absorption by utilizing the method for calculating emissivities outlined in Section III-B. It should be noted that approximate heat-transfer calculations by the method proposed in this report will, in general, become less accurate as the rotational spacing increases. For example, emissivity calculations for HF (rotational spacing $\simeq 40$ cm^{-1}) will be considerably less reliable than for CO (rotational spacing $\simeq 4$ cm^{-1}) at a pressure of a few hundred pounds per square inch, because of inadequate overlapping of the rotational lines of HF. Since considerable uncertainty exists regarding the numerical value of the integrated absorption and of the corresponding effective charge, we shall present here only a tentative compilation of experimental data without extensive emissivity calculations. More reliable values for the integrated absorption should become available as the result of current measurements on diatomic molecules.[23]

A. Effective Band Widths for Diatomic Molecules at Different Temperatures

The results of effective band width calculations for CO have been presented in Tables I and II. Similar data for HI, HBr, HCl, and HF are given in Tables III to X.

It is a simple matter to calculate the ratios

$$\epsilon_F' = \left\{\sum_n \int_{\Delta\nu_{n\to n+1}} \rho(\nu)d\nu\right\} \Big/ \int_0^\infty \rho(\nu)d\nu$$

and

$$\epsilon_0' = \left\{\sum_n \int_{\Delta\nu_{n\to n+2}} \rho(\nu)d\nu\right\} \Big/ \int_0^\infty \rho(\nu)d\nu$$

by use of the effective band widths listed in Tables I to X. The results of these calculations for CO have been plotted in Fig. 2 and were discussed earlier in this paper. Limiting emissivities for the hydrogen halides obtained in this manner are shown in Figs. 3 to 6.

B. Effective Charges of Diatomic Molecules

The integrated absorption for the fundamental and the first overtone of CO was measured by a novel indirect method by Matheson in 1932.[24] He reported the values

$\alpha_{01} = 1.182\times 10^{13}$ (cm-atmos.)$^{-1}$ sec.$^{-1}$ at 300°K,
$\alpha_{02} = 1.54\times 10^{11}$ (cm-atmos.)$^{-1}$ sec.$^{-1}$ at 300°K.

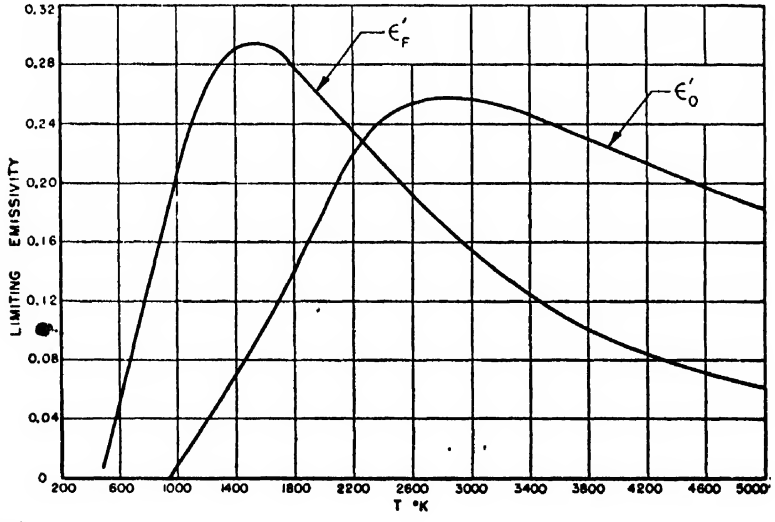


FIG. 6. Limiting emissivity of HF as a function of temperature.

[23] S. S. Penner and D. Weber (to be published).
[24] L. A. Matheson, Phys. Rev. 40, 813 (1932).

From the value of α_{01} it is seen that the effective charge of CO is $\epsilon_F = 3.9 \times 10^{-10}$ e.s.u.

The effective charges ϵ_F for three of the hydrogen halides have been determined by Bartholomé.[25] This work has been criticized by Kemble[26] because of lack of adequate spectral resolution used in making absorption measurements. In this connection it should also be noted that the value for the integrated absorption of HCl reported by Bartholomé is considerably smaller than the corresponding value observed by Bourgin some years earlier,[27] which was later shown to be in agreement with the value of the effective charge calculated from infrared dispersion measurements.[28]

For HI, HBr, and HCl Bartholomé[25] reported effective charges of $0.033e$, $0.075e$, and $0.086e$, respectively, where e is the charge of an electron. In order to estimate ϵ_F for HF, it was decided to extrapolate the values for the other hydrogen halides by plotting ϵ_F/e as a function of the electric moment μ' divided by the normal interatomic distance r_0 of the hydrogen halides.[29] The electric moment μ' for monomeric HF has been reported as 1.91×10^{-18} e.s.u.[30] A plot of ϵ_F/e as a function of μ'/r_0 was found to lead to the result $\epsilon_F \simeq 0.11e$ for HF. According to the measurements of Bourgin[27] and Rollefson and Rollefson[28] it appears likely that these values for the effective charges are too small. As an approximate correction, multiplication by the factor (1/0.42) of all the effective charge values of the hydrogen halides is suggested. The correction factor (1/0.42) represents the ratio of the effective charge of HCl reported by Rollefson and Rollefson[28] to the effective charge of HCl observed by Bartholomé.[25]

Reliable experimental values of the integrated absorption for the first overtone for the hydrogen halides are not available. For this reason it is suggested to calculate $\bar{k}_0$ directly by use of a theoretical relation derived by Rosenthal[31] which can be shown to lead to the relation

$$\frac{k_0}{\bar{k}_F} = \frac{x_e(1-5x_e)}{(1-3x_e)^2}, \tag{24}$$

where x_e is the anharmonicity constant for the molecule under discussion. Equation (24) gives a lower limit for the average absorption coefficient $\bar{k}_0$ since the derivation of this relation includes only contributions to the intensity of the first overtone resulting from the first term in the expansion of the electric moment in terms of internuclear distance. The calculated lower limit for the intensity of the overtones represents a reasonably good approximation to the true intensity in some cases. This assertion has been verified for HCl by comparison with measured intensity ratios.[31] The ratio $x_e(1-5x_e)/(1-3x_e)^2$ is seen to be somewhat smaller than the integrated intensity ratio of overtone to fundamental for CO, using Matheson's data.[24]

C. Application to Emissivity Calculations

The results listed in Tables I to X, together with the quantitative absorption measurements summarized in Section IV-B, provide the data necessary for approximate emissivity calculations at elevated total pressures. Because of the uncertainty in the quantitative absorption measurements, we shall not present extensive emissivity tables at this time but rather illustrate the method of calculation by a representative example.

It is desired to determine the emissivity of HF at 3000°K when the gas is uniformly distributed through a radiation path length l of 10 cm and the partial pressure of undissociation HF is 15 atm.[32] The total pressure is 25 atm.

From the data given in Section IV-B the average absorption coefficients are found to be $\bar{k}_F = 2.41 \times 10^{-3} \times (1/0.42)^2 = 13.7 \times 10^{-3}$ (cm×atmos.)$^{-1}$ and $\bar{k}_0 = 4.64 \times 10^{-5} \times (1/0.42)^2 = 26.3 \times 10^{-5}$ (cm×atmos.)$^{-1}$, respectively. Using the limiting emissivity plots of Fig. 6 we find for the upper limits of the emissivities of the fundamental and the first overtone the values 0.155 and 0.256, respectively. Hence the emissivity ϵ' is

$$\epsilon' = [1-\exp(-13.7\times10^{-3}\times15\times10)]0.155 + [1-\exp(-26.3\times10^{-5}\times15\times10)]\times0.256 = 0.144.$$

The author takes pleasure in expressing his appreciation for many stimulating discussions to Dr. Martin Summerfield, who initiated the program of the calculation of gaseous emissivities from spectroscopic data. The numerical calculations were performed by the computing group of the Jet Propulsion Laboratory under the supervision of Mrs. Macie Roberts.

[25] E. Bartholomé, Zeits. f. Physik **B23**, 131 (1933).
[26] E. C. Kemble, J. Chem. Phys. **3**, 316 (1935).
[27] D. G. Bourgin, Phys. Rev. **29**, 794 (1927).
[28] R. Rollefson and A. H. Rollefson, Phys. Rev. **48**, 779 (1935).
[29] This extrapolation of observed values of ϵ_F for HCl, HBr, and HI to yield the effective charge ϵ_F for HF is based on the relation $\epsilon_F = (d\mu'/dr)_{r=r_0} \simeq 3\mu'/r_0$ which, although apparently not valid for the hydrogen halides, may give a reasonable relation for the change of ϵ_F with μ'/r_0. The effective charge of HF will be measured in the near future. The effect of association of HF on emissivity has been neglected since association would not be expected to be of importance at elevated temperatures.
[30] N. B. Hannay and C. P. Smyth, J. Am. Chem. Soc. **68**, 171 (1946).
[31] J. E. Rosenthal, Proc. Nat. Acad. Sci. **21**, 281 (1935).
[32] It should be noted that p always refers to the partial pressure of the undissociated emitter since the atomic species H, C, O, I, Br, Cl, and F do not radiate in the infra-red and therefore do not make an appreciable contribution to radiant heat transfer at temperatures below about 5000°K.

X-Ray Measurement of Long Range Order in β-Brass*

D. CHIPMAN AND B. E. WARREN
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received February 6, 1950)

The long range order in β-Brass has been determined by measuring the integrated intensity of the (100) superstructure reflection from a single crystal held at various temperatures. With a single crystal, and CuKα-radiation obtained by a balanced Ni-Co filter, there is ample intensity for measurement with a Geiger counter spectrometer. The measured long range order parameter S is in satisfactory agreement with the theoretical predictions.

I. INTRODUCTION

THE β-Brass phase exists at about the composition CuZn. There is a critical temperature of approximately 465°C, above which the structure is disordered body-centered cubic, and below which it is an ordered CsCl structure. In a perfectly ordered AB structure, the A atoms are at α-sites and the B atoms at β-sites. In a structure with partial long range order, if r_α represents the fraction of α sites rightly occupied by an A atom, the Bragg and Williams[1] long range order parameter is given by $S=2r_\alpha-1$. For perfect order $S=1.0$ and for complete disorder $S=0.0$. For an AB structure with partial long range order, the structure factor is given by

$$F=[f_B r_\beta+f_A(1-r_\beta)]+[f_A r_\alpha+f_B(1-r_\alpha)]\times\exp[\pi i(h+k+l)].$$

For the superstructure reflections $h+k+l=$ odd

$$F=(f_B-f_A)(2r_\alpha-1)=(f_B-f_A)S. \qquad (1)$$

The integrated intensity is proportional to F^2, and hence a measurement of the integrated intensity of a superstructure reflection gives directly the square of the long range order parameter.

Since $(f_{Zn}-f_{Cu})$ is small, superstructure lines on a powder pattern of β-Brass are extremely weak. The difference in scattering factors has been increased by the use of ZnKα-radiation[2] and it can be further enhanced by the use of crystal monochromated CuKβ-radiation. However, by using a single crystal of β-Brass, the (100) reflection is strong enough to be readily measured. In general, with crystal monochromated radiation there will be a half-λ from the continuous spectrum reflected from (200) of the sample and superimposed on the (100) peak. This can be eliminated by using (111) of fluorite as the monochromator. We have used CuKα-radiation with a balanced Ni-Co filter and the tube run at 15 kv constant potential. This eliminated half-λ-difficulties and gave ample intensity for the measurements.

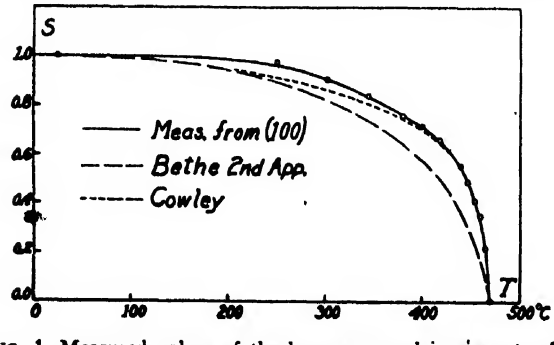


FIG. 1. Measured values of the long range order parameter for β-Brass compared with theoretical values.

II. MEASUREMENT OF LONG RANGE ORDER IN β-BRASS

The single crystal of β-Brass was cut parallel to a (100) face and polished and etched. It was mounted in a small furnace provided with a Ni foil window for the primary and diffracted beams. A slow stream of nitrogen was used to prevent oxidation. The temperature was measured by a thermocouple fastened to the back of the crystal. The crystal was mounted at the center of a Geiger counter spectrometer. With a wide counter slit and fixed position of the counter, rocking curves were made plotting intensity as a function of crystal angle.

The long range order at room temperature was measured by comparing the integrated intensities of (100) β-Brass and (600) NaCl.

$$A_{CuZn}\left[\frac{F^2(L\cdot P)e^{-2M}}{\mu v_a^2}\right]_{NaCl}=A_{NaCl}\left[\frac{F^2(L\cdot P)e^{-2M}}{\mu v_a^2}\right]_{CuZn}, \qquad (2)$$

where A is the measured rocking curve area, F is the structure factor, $(L.P.)$ is the Lorentz-polarization factor, e^{-2M} is the temperature factor, μ is the linear absorption factor, and v_a is the unit cell volume. The Hönl correction[3] was applied to the scattering factors. From (2) we obtain F for β-Brass in absolute units, and using (1) we get directly the long-range order parameter S. The value obtained was $S=1.08$. This value has been interpreted as meaning that within the

* Research sponsored by the ONR under Contract N5-ori-07832.
[1] W. L. Bragg and E. J. Williams, Proc. Roy. Soc. 145, 699 (1934).
[2] F. W. Jones and C. Sykes, Proc. Roy. Soc. 161, 440 (1937).
[3] R. W. James, *The Optical Principles of the Diffraction of X-Rays* (G. Bell and Sons, London, 1948), p. 608.

experimental error the long range order in β-Brass at room temperature is essentially perfect.

To measure long range order as a function of temperature, the crystal of β-Brass was held at a series of temperatures up to and above $T_c=465°C$. A rocking curve was made at each temperature. The rocking curve areas, corrected for a temperature factor e^{-2M}, are proportional to S^2. Normalizing the areas to make $S=1.00$ at room temperature, the values of S shown in Fig. 1 were obtained. The measured values of S are compared with the theoretical curves of Bethe[4] and Cowley.[5] The Cowley curve is a plot of the equation

$$\ln(1+S^2/1-S^2)=2S^2(T_c/T), \qquad (3)$$

using $T_c=738°K$. There is qualitative agreement with either curve. The quantitative agreement with the Cowley curve is probably within the experimental error.

III. ANTIPHASE DOMAINS IN β-BRASS

If antiphase domains in an ordered structure are small enough, they show up by a broadening of the superstructure lines. Although antiphase domains are to be expected in ordered Cu_3Au, the structure of ordered CuZn is such that they are not to be expected. However, it was thought worth while making a set of line breadth measurements to test the point. The counter slit was made very narrow and the counter set at a series of positions to plot out the diffraction peak. At each position of the counter, the crystal was turned at constant velocity through the reflecting range, and the total count recorded. The peak breadth obtained in this way is independent of any mosaic structure in the crystal.

Two samples were prepared; one cooled from above T_c to room temperature in five hours, and the other quenched from 500°C into brine. Both samples were then polished and etched, and peaks plotted out for the (100) reflection. Within experimental error the two breadths were the same--about 11 min. in 2θ. Although a measurement of the breadth of the (100) peak is not a sensitive test for antiphase domains, the results indicate that there is no prominent antiphase structure in β-Brass.

IV. EFFECT OF ANTIPHASE DOMAINS ON THE INTEGRATED INTENSITY OF A SUPERSTRUCTURE LINE

It is important in the determination of the long range order parameter that the integrated intensity of a superstructure reflection is independent of the existence of antiphase domains.

Imagine the ordered crystal to be divided into blocks N_1a_1, N_2a_2, N_3a_3, where N_1, N_2, N_3 are numbers of the order 10–20. Cells within a block are numbered $m_1m_2m_3$ and the blocks are numbered $n_1n_2n_3$. To allow for any system of antiphase domains or displacements within the crystal we assign an arbitrary phase $\varphi_{n_1n_2n_3}$ to the amplitude from each block. The amplitude of diffracted radiation from the crystal is then

$$\begin{aligned}\text{Amp}=F\textstyle\sum_m \exp[(2\pi i/\lambda)(s-s_0)\cdot(m_1a_1+m_2a_2+m_3a_3)]\\ \times\sum_n \exp[i\varphi_n]\exp[(2\pi i/\lambda)(s-s_0)\\ \cdot(n_1N_1a_1+n_2N_2a_2+n_3N_3a_3)],\end{aligned}$$

where F is the structure factor per cell and s, s_0 are unit vectors giving the directions of the diffracted and primary beams. Let

$$s-s_0=\lambda(h_1b_1+h_2b_2+h_3b_3),$$

where $b_1b_2b_3$ are the reciprocal vectors and $h_1h_2h_3$ are continuous variables. The intensity from the crystal is then

$$\begin{aligned}I(h_1h_2h_3)=F^2\textstyle\sum_m\sum_{m'}\exp[2\pi i\{h_1(m_1-m_1')\\ +h_2(m_2-m_2')+h_3(m_3-m_3')\}]\\ \times\{N_B+\sum_{\substack{n\ n'\\ n\neq n'}}{}'\exp[i(\varphi_n-\varphi_{n'})]\\ \times\exp[2\pi i\{h_1N_1(n_1-n_1')+h_2N_2(n_2-n_2')\\ +h_3N_3(n_3-n_3')\}]\} \qquad (4)\end{aligned}$$

where N_B is the number of blocks in the crystal.

The measured integrated intensity of a reflection is proportional to an integral of the interference function throughout a sufficient region in reciprocal space about the point hkl to include everything contributing to the hkl reflection.

$$P=K\int\int\int I(h_1h_2h_3)dh_1dh_2dh_3. \qquad (5)$$

Substituting (4) in (5) and considering a term $n\neq n'$

$$\begin{aligned}\int_{h-\frac{1}{2}}^{h+\frac{1}{2}}\exp[2\pi ih_1\{(m_1-m_1')+N_1(n_1-n_1')\}]dh_1\\ =\exp[2\pi ih\{(m_1-m_1')+N_1(n_1-n_1')\}]\\ \times\frac{\sin\pi\{(m_1-m_1')+N_1(n_1-n_1')\}}{\pi\{(m_1-m_1')+N_1(n_1-n_1')\}}.\end{aligned}$$

The integrals are zero for all terms where $n\neq n'$ and

$$\begin{aligned}P=KF^2N_B\int\int\int\textstyle\sum_m\sum_{m'}\exp[2\pi i\{h_1(m_1-m_1')\\ +h_2(m_2-m_2')+h_3(m_3-m_3')\}]dh_1dh_2dh_3.\end{aligned}$$

The measured integrated intensity is the number of blocks N_B times the integrated intensity from each block and independent of the arbitrary phases φ_n. Hence a measurement of the long range order parameter S from the integrated intensity of a superstructure reflection is independent of whether or not antiphase domains exist in the ordered structure.

[4] H. Bethe, Proc. Roy. Soc. **150**, 552 (1935).
[5] J. M. Cowley, Phys. Rev. **77**, 667–675 (1950).

The Resolving Power of an X-Ray Microscope

EDWARD PRINCE*
General Electric Research Laboratory
(Received February 10, 1950)

A formula for determining the resolving power of an x-ray microscope built according to the design suggested by Kirkpatrick and Baez is derived, and an estimate of that resolving power given. The optimum resolving power is such as to resolve points separated by a distance of the order of 2000A.

INTRODUCTION

PAUL Kirkpatrick and A. V. Baez[1] have described a method of forming optical images with x-rays, making use of the total reflection at grazing incidence. The x-rays are reflected from two cylindrical mirrors in such a way that one mirror brings all horizontal lines to a focus, and the other brings all vertical lines to a focus. On the basis of diffraction theory, Kirkpatrick and Baez set the limit of resolution of such a system at about 70A, assuming perfect geometrical focusing and unit magnification. Since perfect focusing with mirrors is a condition which may be achieved only for infinitesimally small objects, these assumptions are not directly applicable to the analysis of an actual x-ray microscope built according to this design, and it is, therefore, desirable to make an estimate of the optimum conditions in a practical instrument, taking into consideration both the geometrical and the diffraction effects at the same time.

MATHEMATICAL ANALYSIS

In order to evaluate the combined effects of geometrical aberration and diffraction, let us first derive an applicable formula for each in turn. The effect of diffraction can be determined easily from the usual

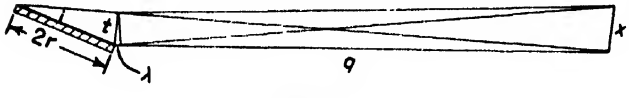


FIG. 1.

formula (see Fig. 1):

$$x=(\lambda q)/t \tag{1}$$

where x is the separation of two resolved spots in the image; t is the width of the aperture; q is the distance from the aperture to the image; and λ is the wavelength. Let y be the resolution, p the distance from the object to the aperture, and $2r$ the width of the mirror. The resolution is then:

$$y=(xp)/q \quad \text{or} \quad y=(\lambda p)/t. \tag{2}$$

If i is the critical glancing angle, and is small, so that $\sin i=i$, then the resolution due to diffraction is given by:

$$t=2ri, \quad y=(\lambda p/2ri). \tag{3}$$

It is now necessary to find an expression for the resolution in terms of the geometrical aberrations. This is difficult, because it is dependent on the distribution of intensity around the focus, but the distance by which the extreme ray misses the focus gives a sufficient condition which probably does not differ from the necessary condition by more than a factor of 2 or 3. Kirkpatrick and Baez evaluate this quantity for a circular section:

$$y=\tfrac{3}{2}(r^2/R) \tag{4}$$

where R is the radius of curvature and $R\gg r$. A minimum value of y may be obtained by equating expressions 3 and 4, giving:

$$\frac{\lambda p}{2ri}=\frac{3r^2}{2R}, \quad r=\left(\frac{\lambda pR}{3i}\right)^{\frac{1}{3}} \tag{5}$$

or

$$y=\left(\frac{3\lambda^2 p^2}{8Ri^2}\right)^{\frac{1}{3}}. \tag{6}$$

CONCLUSION

If $p=10$ cm $R=20$ m and i is the value used by Kirkpatrick and Baez, $i=7\times10^{6}\lambda$, then y is equal to about 3400A. By the use of different curves and with skillful operation, it is probable that this figure can be reduced by a factor of 2, and possibly by as much as a factor of 4, but it is highly doubtful that the limit can be reduced much more than that in a practical instrument with reasonable magnification. This makes the x-ray microscope potentially an instrument with resolution between that of the ordinary optical microscope and that of the electron microscope, with an advantage of high penetrating power.

The author wishes to express his thanks to Dr. David Harker and Miss C. M. Lucht, of the General Electric Research Laboratory, who suggested this investigation.

* Present address: Cavendish Laboratory, Cambridge University, Cambridge, England.
[1] P. Kirkpatrick and A. V. Baez, J. Opt. Soc. 38, 766 (1948).

Metallic Evaporation and the Diameter of Tobacco Mosaic Virus with the Electron Microscope*

H. KAHLER AND B. J. LLOYD, JR.
National Cancer Institute, Bethesda, Maryland
(Received March 13, 1950)

Tobacco mosaic virus has been used as a known test object for the intercomparison of several methods of small particle measurements, all involving metallic evaporation. Various characteristics of these methods, which determine their accuracy, were investigated. Measurement of groups and the derivation of a mean size therefrom is primarily limited in accuracy only by the exactness of juxtaposition of the successive particles in the group. The measurement of the images of single particles is subject to the relatively large error associated with the uncertainty in the relationship between the true particle edge and the image of the edge which has received a coating of metal. The measurement of shadow lengths reduces this error by the factor $2\tan\alpha/2$. A symmetrical shadow method is described which reduces the error of substrate tilt, and is insensitive to amount of metal deposited. An average over the four best methods gave a value 150.0±3.6A agreeing closely with the x-ray determination.

INTRODUCTION

ONE of the most important problems in electron microscopy is the measurement of the size of small particles. The percentage errors which arise in such measurements obviously should increase as the size of the particle diminishes down to the resolving power of the instrument. The object of the present investigation was to intercompare several methods of measurement in regards to general accuracy and reliability. Tobacco mosaic virus, TMV, is one of the few small objects which have been extensively studied for size (diameter) and its use as a test object would appear to be satisfactory for the present study. The methods studied all involved the evaporation of metallic films upon the specimen; it was, therefore, necessary to make some supplemental studies upon the relation of metal film to TMV particle.

While this study was under way an inter-laboratory comparison of a much larger particle (polystyrene latex Dow No. 580) was initiated by Gerould[1] following the work of Williams and Backus. Agreement to ±2 percent was obtained by 12 out of 16 laboratories for the mean of the particle size, the value reported by us being 2600±65A, which was within 0.4 percent of the inter-laboratory mean. In certain phases of our work this material was therefore used for determination of magnification power.

METHOD

The preparation of TMV was put on Formvar films (the Formvar having been cleared of large aggregates by a one hour centrifugation at 20,000×g). The different arrangements of components in the vacuum evaporator is illustrated in Fig. 1. Single filament evaporation takes place with the filament at Y_1, and double evaporation with filaments at Y_1 and Y_2. The angle α, of evaporation, is 90 degrees with the filament at the position Y (vertical evaporation). When the long axis of the TMV is in the direction OX, it is parallel to the evaporation beams from Y_1 and Y_2, and when in the direction OZ it is perpendicular to the evaporation beam.

The electron microscope was used without an objective aperture. The microscope calibration was done with a 20,000 line/inch diffraction grating replica, coated with chromium, and in some cases with polystyrene latex (Dow No. 580). Low magnification photos (×8000) were made with the unbiased gun and large angle of illumination while high resolving power photos at ×23,000 were taken with the biased gun, compensated objective and an additional fine focusing potentiometer.

Method I. Group Measurement

It is possible to obtain the metal coated specimen in the form of a horizontal linear array of particles or monomolecular film,[2] which is in some respects a one-dimensional crystal. Measurement across a number of the particles, excluding the extremes of the array, gives

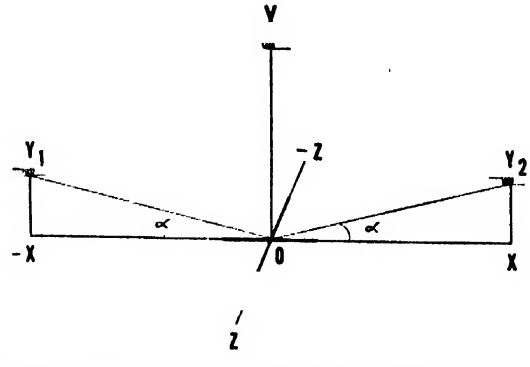


FIG. 1. Sketch of set-up in vacuum evaporator. Filaments at Y_1 for single evaporation, at Y_1 and Y_2 for double evaporation and at Y for vertical evaporation. Long axis of TMV in direction OX, evaporation parallel to axis; TMV in direction OZ, evaporation perpendicular to TMV axis.

* This paper, in part, was presented at the Toronto meeting of the Electron Microscope Society of America, September 11, 1948.

[1] C. H. Gerould, J. App. Phys. **21**, 71 (1950); **21**, 183 (1950).

[2] R. C. Williams and R. W. G. Wyckoff, Science **101**, 594 (1945).

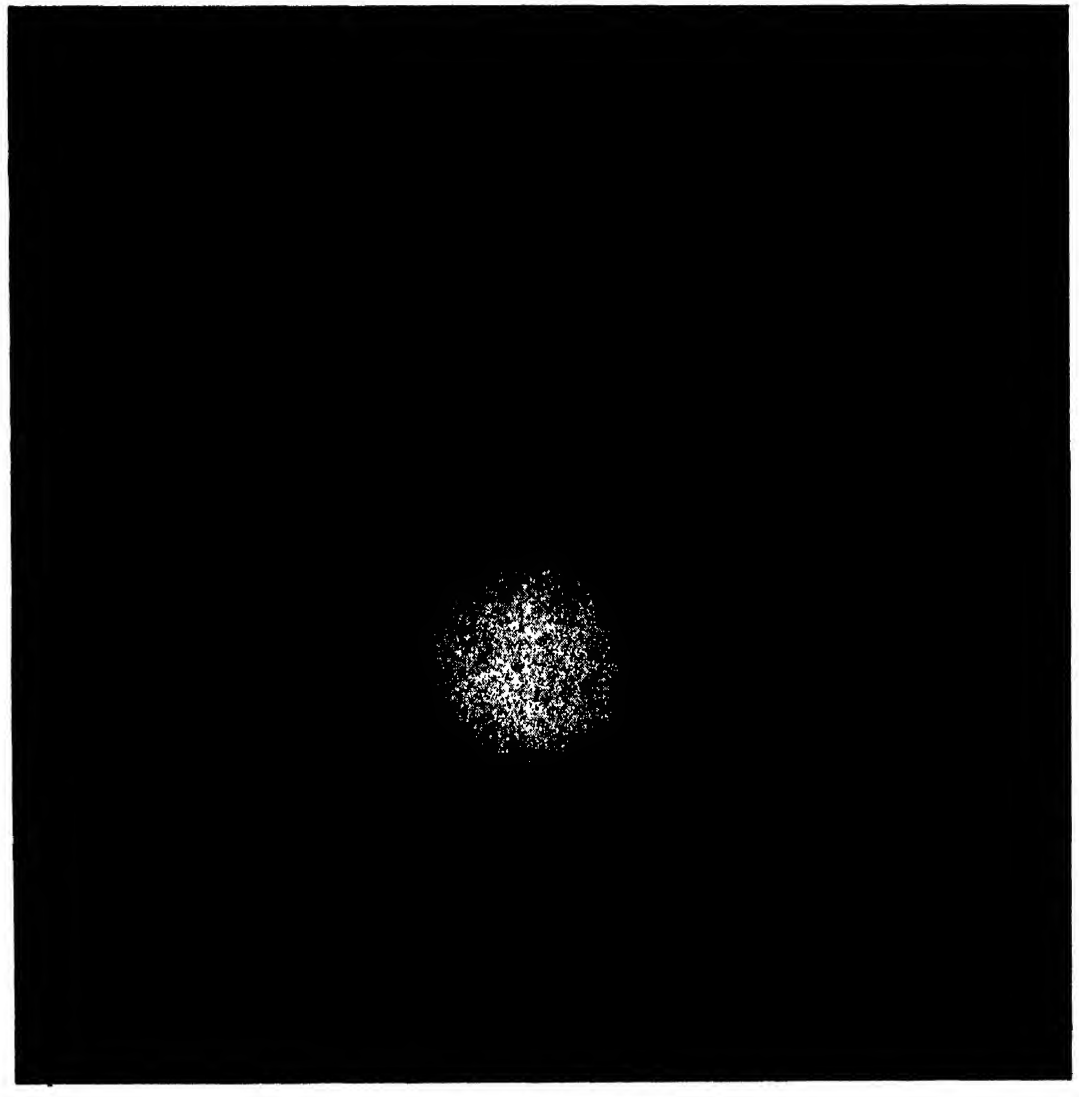


FIG. 2. Light Cr double evaporation 1.5 mg at angle 2 cm to 10 cm each filament. Circle in center of photograph is polystyrene latex, diameter 2600A (unshadowed), which is underfocused as shown by diffraction ring; compensated objective, biased gun. Arrows indicate direction of chromium evaporation beams. Print of positive.

an average particle spacing which is independent of the metal coating and boundary diffraction effects. The chief uncertainty in this procedure is in the unknown juxtaposition of consecutive particles. It was shown by Bernal and Fankuchen[3] that the minimum particle spacing in a TMV gel was 175A, with a hexagonal close packed two-dimensional lattice. When this gel was air dried to a hexagonal crystal the spacing shrank to 152A decreasing to 150A upon vacuum drying over P_2O_5 with a high degree of uniformity in size. These authors postulated a water shell on the particles about 10A thick, preventing a closer approach than 175A as long as any free water was present. From this it may be concluded that TMV particles deposited from the aqueous film onto the Formvar screen have a minimum spacing of 175A while still in the wet state. For such a system as the dried monomolecular TMV on the Formvar substrate, the x-ray data gives no direct evidence as to the final packing dimension, though it could be argued that the spacing should be somewhere between 150 and 175A.

Experimentally,[4] a value of 152 without an estimate of deviation has been reported, and a value 130±6A was assigned to a related cucumber virus where the 6A refers to deviation between group means σ_M rather than individual variation. We have found that the mean for different groups of 8 to 20 TMV particles range in size from 152 to 160A with a standard deviation σ_M between groups of 3A and a mean of 155A. The variation in packing and measurement of individual particles

[3] J. C. Bernal and I. Fankuchen, J. Gen. Physiol. 25, III (1941).

[4] C. A. Knight and G. Oster, Arch. Biochem. 15, 289 (1947).

σ deduced from groups of 10 would obviously be larger as $\sigma = (n-1)^{\frac{1}{2}}\sigma_M = 9$A where n is the number of particles per group. From this it would appear that those groups of particles which have the smallest spacing or are packed the closest give nearest to the x-ray measurement for a completely desiccated particle. It also indicates that desiccation on the screen is accompanied by a shrinkage of the group spacing from 175A to a value close to 150A.

A variation of this procedure namely a replication of crystal surfaces should give an even closer correspondence with Bernal and Fankuchen's results.

It appears that the group method gives a mean value whose error is only slightly larger than the usual calibration errors of the electron microscope.

Method II. Direct Measurement on Doubly Coated Particles

This method consists in evaporating a thin coating of metal, from two oppositely placed filaments, upon opposite edges of the particles, and measuring the distance across the highly magnified image of particles having their long axis perpendicular (*OZ*) Fig. 1 to the metal beam direction.

For the direct measurement of particle diameter, the electron microscope was fitted with a compensated objective to reduce astigmatism[5] and to increase resolving power (estimated to be 20A).

Using the double coating method it was found possible to decrease the amount of evaporated chromium to the point where shadows were scarcely visible but contrast of the particle edges was sufficient for measurements, (Fig. 2). In this photo the apparent widening of the particles is evident as the particles approach a perpendicular direction to the evaporation beam. This widening is more rapid than would be expected on the basis of the amount of metal deposited on the particle. Actually, the particles parallel to the evaporation beams are narrower than the calculated value. In Fig. 3 there is shown a focal series for two particles, one a long aggregated virus rod nearly perpendicular to the direction of the evaporation beams (arrows) and a shorter rod nearly parallel to the beam direction. Figure 3 is reproduced as a print of a negative since previous authors have used such prints in discussing diffraction patterns. As an aid to the estimation of the in-focus position, the diffraction fringes on a polystyrene latex sphere placed on the screen after shadowing were noted (Fig. 2). The magnitude of the out-of-focus position was estimated from the data obtained for displacement of the specimen *versus* objective lens currents after the method of Hillier and Ramberg.[5]

This series is characterized by two minima (black) near each edge of the rods for the overfocused positions and a single minimum near the edge for the underfocused images. The measurement across the diameter

[5] J. Hillier and E. G. Ramberg, J. App. Phys. 18, 48 (1947).

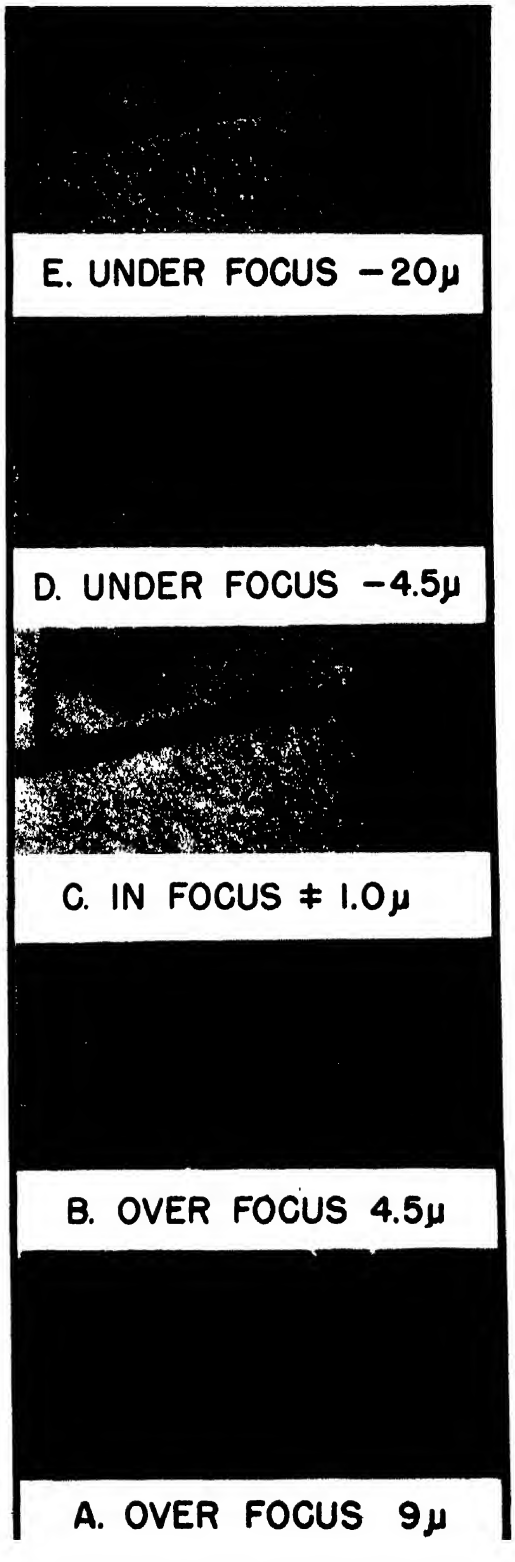


FIG. 3. Focal series of TMV double coated with 2×5.5 mg Cr $\tan\alpha = 2/12$ biased gun, compensated objective. In-focus position estimated with aid of diffraction pattern of polystyrene latex spheres added to the specimen screen after shadowing (not shown). Direction of evaporation beam shown by arrows. Print of negative.

of the particles between fringe centers in the three upper pictures (*c*, *d*, *e*) is approximately the same (to within ±10A); in the lower pictures (*a*, *b*) a similar value is obtained when the measurement is made between the mid-point, (maximum), of each pair of fringes. Thus, over a focal range of 29μ, an invariant measurement (to within ±10A) is obtained between selected points of the edge fringes.

The diffraction minima of the metallic coating cannot be exactly defined in position with respect to the original uncoated particle edge. Part of this uncertainty lies in the variable difference between effective thickness of the metal and its calculated thickness. When a light coating of chromium is deposited on the particle edges in a perpendicular direction, the intensity minima of the slightly underfocused images lie between the middle of the coating and the outer coating edge.

A series of measurements showed that the distance between fringe minima was 10±20A larger than the calculated distance between the centers of the metallic coatings.

For other positions of the TMV such as long axis parallel to the evaporation beam the disposition of metal is no longer principally upon the extreme edges so that frequently the images are too narrow. The same holds for vertical evaporation (filament at *Y* Fig. 1).

Method III. A Combination of I and II

A light symmetrical coating is deposited as in method II, producing intensity minima near the edge of the particles. Then two particles lying side by side are compared with a single particle close by and parallel in order to eliminate the edge error. For the single particle the measured distance between fringe centers reduced by the magnification factor is,

$$X_1 = D + 2e$$

where *e* is the amount by which the fringe deviates from the true edge, *D* is the true diameter.

For the double particle,

$$X_2 = 2D + 2e$$

from which

$$X_2 - X_1 = D.$$

The value obtained by this procedure was 152±11A.

The chief precautions to be observed are equality of shadowing and, as in method II, the use of high magnifications.

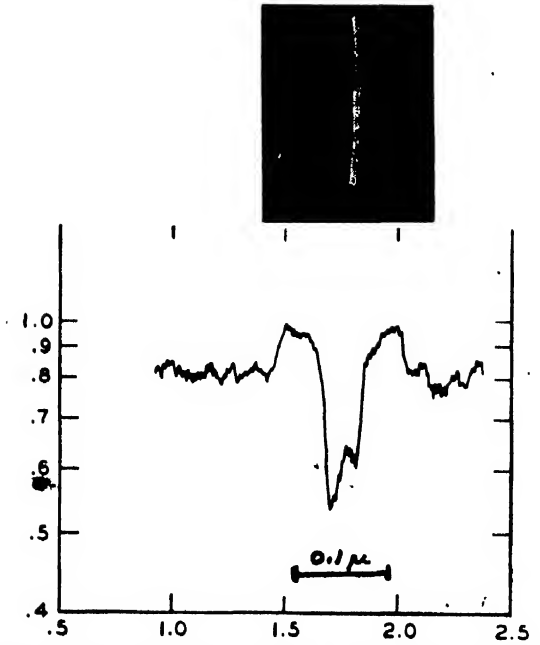


FIG. 4. Microphotometer tracing across a photo of a double shadowed virus particle. Two shadow beams of chromium were unequal, resulting in unequal metallic coatings. Print of a positive. Non-biased gun, low power original.

Method IV. Direct Measurement Following Single Metallic Coating

Coatings of gold of a series of thicknesses were deposited on the TMV by evaporation from a single filament at an angle of 10 degrees. The measurement was made across the particle. This method has all the faults of the double coating method (II) and in addition, the more serious error which asymmetry introduced into the measurement giving values too erratic to quote.

Method V. Double Shadows

The set-up for the production of double shadows is identical with that of the double coating procedure (II), with evaporation sources at Y_1, Y_2. The only difference is in the employment of a larger amount of metal to be evaporated.

A microphotometer tracing (courtesy of W. C. White of these laboratories) of a photograph of a heavily shadowed particle is shown in Fig. 4. In this experiment, more metal was incident on the particle in one direction than in the opposite direction resulting in unequal shadows and a displacement of the central line. Figure 5 illustrates the case where the two metal beams are nearly equal.

One of the sources of error in the single shadow method (VI) described in a following section, is variation in shadow length due to unknown local tilting of the substratum.[6,7] In the double shadow method, this error is considerably diminished, since a tilt in the substrate results in a shortening of one shadow and a lengthening of the other shadow, the two errors partially compensating.

Thus, if the shadowing beam in the evaporator makes an angle α with the horizontal and the substrate is tipped at an angle β with the horizontal, $2X$ being the length of the total shadow on the horizontal plane, the length of the shadow on the tilted plane is

$$B_1 + B_2 = X\left[\frac{\sin\alpha}{\sin(\alpha+\beta)} + \frac{\sin\alpha}{\sin(\alpha-\beta)}\right],$$

[6] F. W. Boswell, British J. Exp. Path. **28**, 253 (1947).
[7] R. C. Williams and R. C. Backus, J. App. Phys. **20**, 98 (1949).

where B_1 is the short shadow, B_2 the long shadow. For $\alpha=10°$ and $\beta=3°$, B_1 is 77 percent of X, B_2 is 142 percent of X while the sum or double shadow length is 110 percent of $2X$.

It is obvious that a simple inspection of a double shadowed photograph will given an estimate of the flatness and angle of the substratum. Such photographs yield fairly consistent values for the particle height

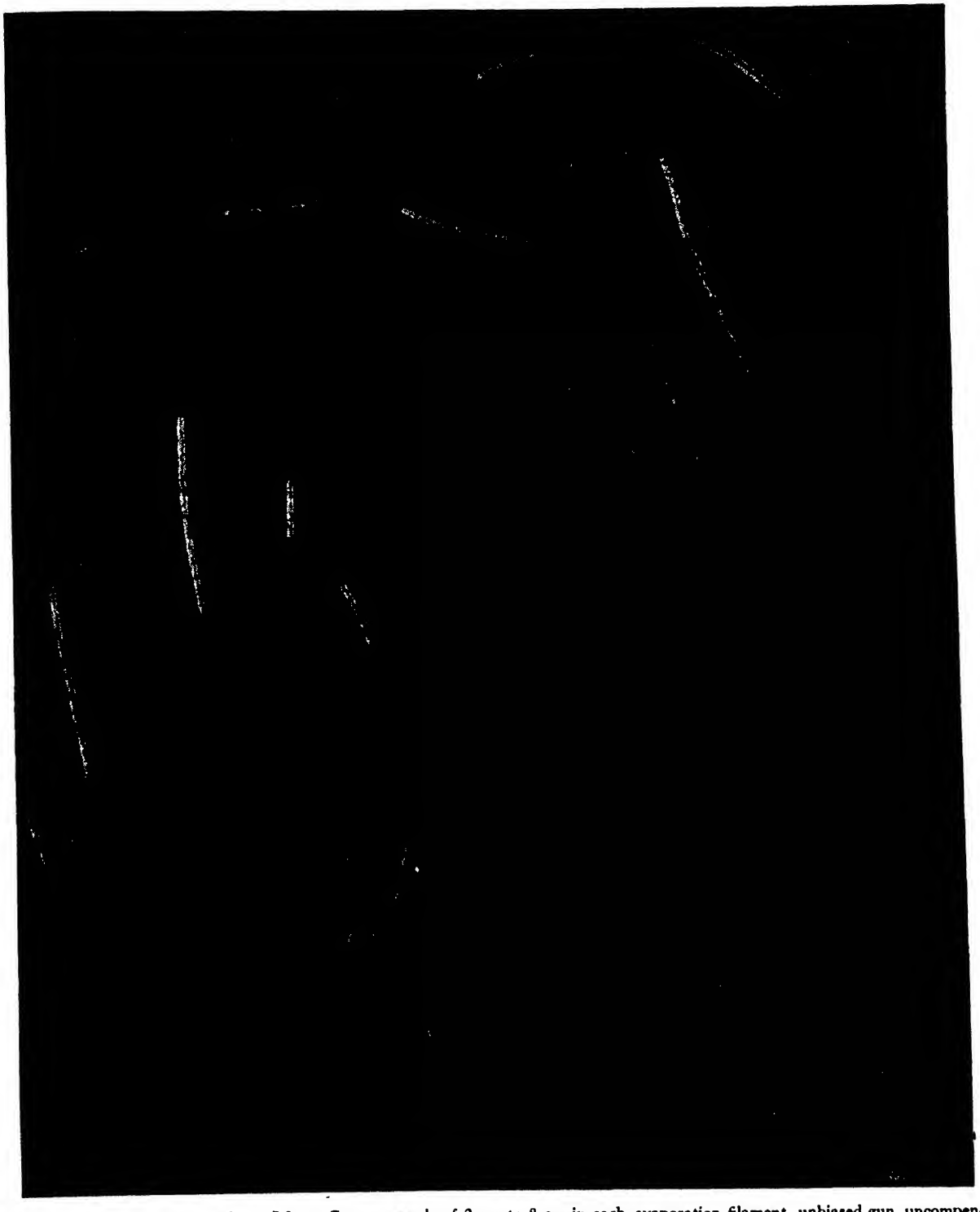


FIG. 5. Heavy Cr double shadows, 7.2 mg Cr at an angle of 2 cm to 8 cm in each evaporation filament, unbiased gun, uncompensated objective. Arrows indicate direction of chromium evaporation beams. Print of a positive.

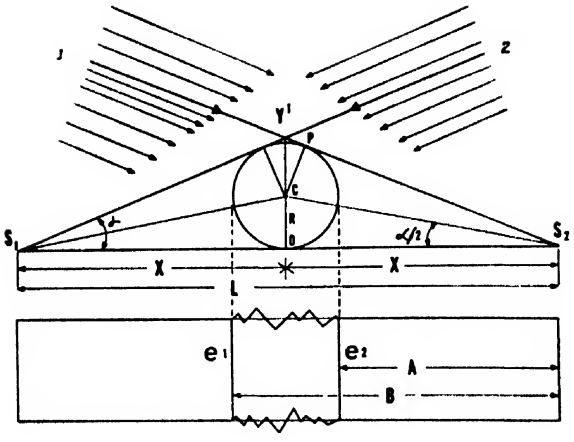


FIG. 6. Diagram of shadowing detail. Single evaporation from filament at 1, shadow measured A or B. Double evaporation from 1 and 2, double shadow length is L.

from $2R=L/M\tan\alpha/2$ where R is the radius of the particle, L is the double shadow length, M the magnification, α is the angle of evaporation. In all shadow methods the use of $2\tan\alpha/2$ instead of $\tan\alpha$ as is ordinarily done is preferable since from Fig. 6 $X\tan\alpha$ gives OY' whereas $2X\tan\alpha/2$ gives $2R$. The error in using $\tan\alpha$ instead of $2\tan\alpha/2$ for 10° is 1 percent, for 20° is 3 percent, for 30° is 8 percent and for 40° is 15 percent. The increase in the length of shadow due to accumulated metal on the top of the particle will be partially compensated by a shortening of the shadow due to an accumulation of metal on the backbround.

The double shadow method has the disadvantage of a low shadow contrast the metal thickness outside the shadow being twice that inside. In order to obtain the maximum shadow contrast, lower magnifications (8000–10,000) have been more satisfactory than photographs taken at 20–25,000 times.

The Formvar screens for these measurements were precoated with a thin layer of evaporated aluminum in order to prevent any sinking of the particles into the Formvar. Then TMV was put on the aluminum surface. In the evaporator the filaments were straight tungsten wires arranged horizontally with Au and Al[8] in the weight ratio 3.5/1 wrapped around the wires. This was done to obtain an accurate measurement of the evaporation angle. Latex No. 580 was then added to the screens for calibration of magnification. Measurements of the photographs of these specimens gave a value of 145.8±8.2A. This is within the expected errors in magnification and determination of exact shadow boundary.

[8] *International Critical Tables*, Vol. II, p. 402.

This method is therefore comparable in accuracy to group method I and method II.

Method VI. Single Shadows

Evaporation of metal from filament I, Fig. 6, gives a cast shadow[9] from which the diameter, $2R$, of the particle may be estimated either by measurement of distances A or B. Since $R=X\tan\alpha/2$, $X=A+R$ and $X=B-R$ by substitution

$$2R=\frac{2A}{M}\frac{\tan\alpha/2}{(I-\tan\alpha/2)}$$

and

$$2R=\frac{2B}{M}\frac{\tan\alpha/2}{(I+\tan\alpha/2)},$$

where M is the magnification. The errors made in estimating A and B are reduced to about 0.2 because of the trigonometric factor with a shadowing angle of 10°. The length B requires a correction for metal thickness at e_1; because of the comparatively heavy coatings required by shadowing methods a large fraction of the image appears opaque. The only recourse is to measure to the outer limit of the image and to subtract the full metal thickness to obtain B. It may be noted that the circle of confusion errors at the extremities of the B measurement tend to cancel out while for the A measurement they are additive.

The procedure which gave the most consistent results was the use of a system of 3 slits interposed between a line filament and the specimen screens 20 cm away in the evaporator in order to define the angle of evaporation, utilizing gold and the low intensity microscope gun. Measurements on 24 views gave for the A measurement 147.6±13A and for the B measurement 152.2 corrected to 146.2±14A. Despite the insignificant difference between these 2 values, the B measurement seems from other series of tests to be the more reliable of the two.

It may be remarked that the single shadow method gives a hybrid measurement, the shadow region corresponding to height and the remainder of the measurement across the particle being width.

SUMMARY OF MEASUREMENTS

An average taken over the better methods should reduce the methodical error. The mean of method (I) 155A, (III) 152A, (V) 146A, (VI) 147A, is 150.0±3.6A. This agrees with the mean value obtained by the x-ray method.

[9] H. O. Müller, Kolloid Zeits. 99, 6 (1942).

The Apparent Size of Objects as Observed in the Electron Microscope*

STANLEY F. KERN AND RAYMOND A. KERN
Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana
(Received March 24, 1950)

The image formed in the electron microscope of shadowed polystyrene latex spheres is larger than the image of unshadowed spheres. This increase in image size and probable causes are discussed. Implications of this phenomena on the study of viruses and other biological specimens are mentioned.

INTRODUCTION

IN our studies with the electron microscope it has been felt that there is a real difference in the image size of small objects depending upon whether they are shadowed or unshadowed. This difference has been thought by some workers to be accounted for by the increased contrast resulting from the deposition of the metal, together with the actual amount of metal deposited. This has not been a satisfying explanation.

After Williams and Backus[1] announced the remarkable uniformity of particle size of the small spheres in Dow Latex No. 580 *G*, Lot No. 3584, some of this material was obtained to use as a rapid internal check of magnification of the electron microscope. A few preliminary experiments indicated that the size of the image of the latex particle is different for shadowed and unshadowed specimens. Since these particles are remarkably uniform for a given method of sample preparation, it was felt that a study of size changes resulting from shadowing could be undertaken using the styrene particles as a standard specimen.

EXPERIMENTAL

Effect of Electron Bombardment

Since the proposed experiments were to involve exposing a specimen to the electron beam for a time longer than usual, and since it is well known that a specimen grows when bombarded,[2-4] it was necessary to ascertain the extent of growth of the Dow Latex particles under conditions of bombardment. In these experiments an early model *B*, RCA electron microscope with an unbiased gun, operating at 60 kv accelerating potential was used. For observation of the specimen a beam current of 250 microamperes was employed with a spot 7 to 10 microns in width, and 15 to 20 microns in length. The findings are summarized as follows:

1. Near the wire of the supporting screen bombardment of the specimen resulted in a rapid growth of the particle. Bombardment for eighteen minutes caused the diameter of the unshadowed balls to increase 570A u. while the shadowed balls had increased 320A u. This may or may not be a significant difference.

2. When observations are confined to the center of the opening the beam does not hit the wire of the supporting screen. Under these conditions bombardment for eighteen minutes resulted in a growth of the particles of about 80 A.° u. In eight minutes no size change could be measured. To more closely approximate the conditions to be expected in later experiments where repeated exposures were to be made, thirty exposures of one unshadowed field were made. No measurable change in size was found until the twenty-fifth exposure. It was concluded, therefore, that ten to fifteen exposures of a field could be made without danger of a measurable particle growth due to electron bombardment, if the field chosen was in the center of an opening.

Effect of Shadowing

In these studies a mixture of platinum and palladium was used as the shadowing metal. Fifteen milligrams of platinum and five milligrams of palladium were evaporated from a 20-mil tungsten filament placed 8.5 cm from the specimen. In all cases except where indicated

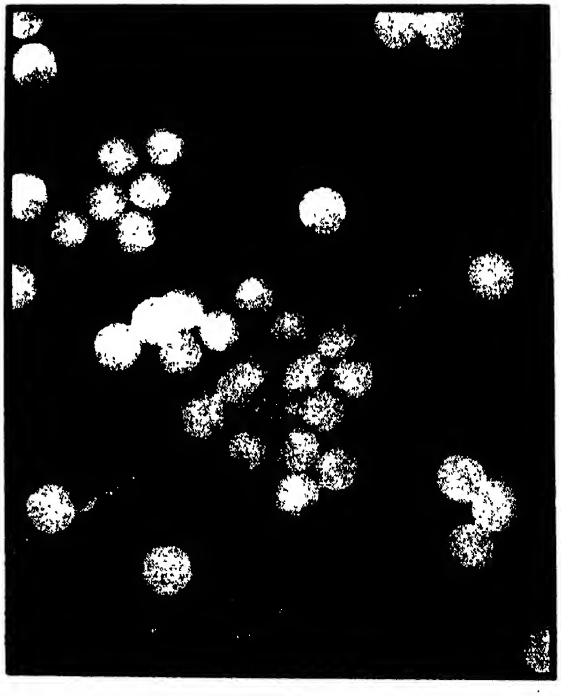


FIG. 1. Shadowed and unshadowed latex particles on the replica of a diffraction grating. 27,000×.

* Presented at the Washington Meeting of the Electron Microscope Society of America October 6–8, 1949.

[1] R. C. Williams and R. C. Backus, J. App. Phys. 20, 224 (1949).
[2] J. H. L. Watson, J. App. Phys. 18, 153 (1947) and 19, 110 (1948).
[3] J. Hillier, J. App. Phys. 19, 226 (1948).
[4] V. E. Cosslett, J. App. Phys. 18, 844 (1947).

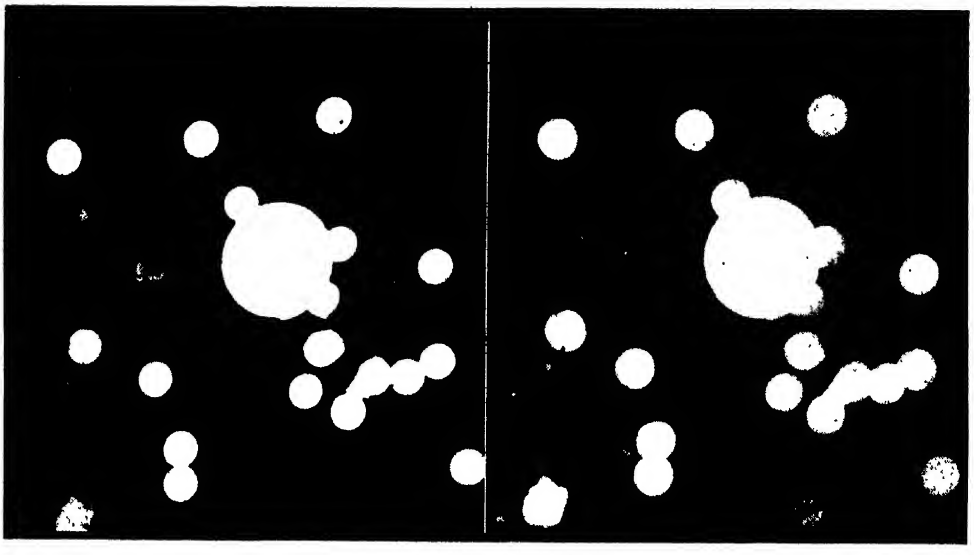


FIG. 2. (a) Latex particles before shadowing. (b) Same field after shadowing. 21,700×.

the angle of shadowing was 10°. A new filament was used for each experiment. The filament with shadowing metal in place was weighed before and after shadowing. The weight loss of the filament and metal was less than the combined weight of the platinum and palladium. It can be safely estimated, therefore, that a layer of metal 10A u. or less was deposited on a flat portion of the field.

Figure 1 is an electron micrograph showing shadowed and unshadowed polystyrene spheres in the same field. A small drop of the latex was allowed to dry on a replica of a diffraction grating and the sample shadowed. Then a second drop of latex was applied to the screen and allowed to dry. The shadowed balls are easily differentiated from the unshadowed balls and their size difference measured. The unshadowed balls measure 2590A u. ±50A u., whereas the shadowed balls measured 2840A u. ±50A u.

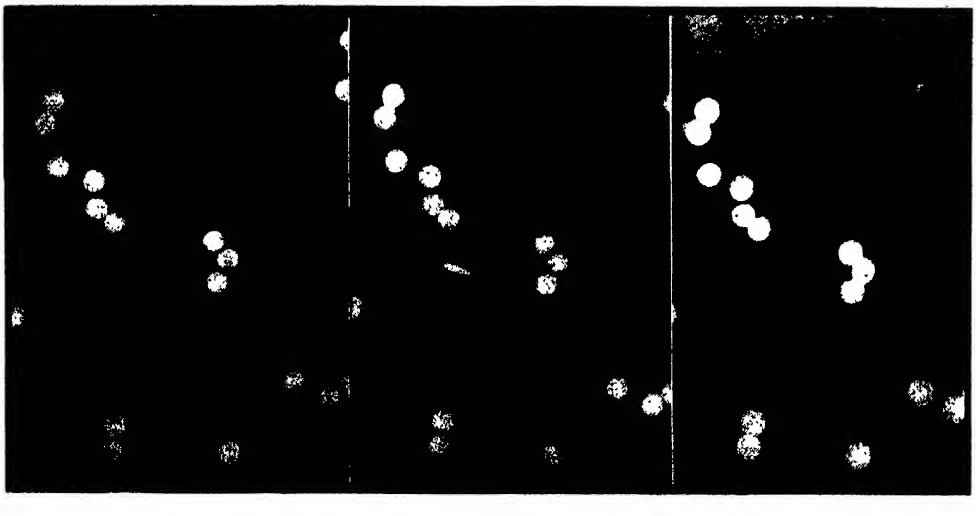
To determine the change of diameter on an individual sphere, micrographs were made of the same field before and after shadowing, Fig. 2. Again the size of the average sphere before shadowing was 2590A u. and after shadowing was 2840A u.

Measurements were made on only those particles which were not in contact with another particle. All measurements on shadowed spheres were made perpendicular to the direction of the shadow.

DISCUSSION

These results show that the image size changes when the specimen is shadowed and that the amount of change is relatively constant. At this time no definite explanation for the observed facts is offered. However, the following have suggested themselves.

One possible explanation is the deposition of extraneous materials during shadowing and electron bombardment. This possibility appears to be ruled out by our



(a) (b) (c)

FIG. 3. (a) Unshadowed latex particles; (b) Same field vertically shadowed; (c) Same field after oblique shadowing. 14,250×.

studies on the effect of electron bombardment on the size of the polystyrene spheres, and, secondly, by the fact that shadowing at a vertical angle does not cause a comparable change in particle size. This experiment will be discussed later. The sharp image obtained also indicates that this is not an over-all deposition of a scattering material.

A second possibility which was considered was the effect of contrast differences resulting from the shadowing. It was felt that possibly the addition of the highly absorbing metal film would materially affect the position of greatest contrast at the edge of the image and would, as a consequence, change its apparent size. Consideration of such a phenomenon shows that this effect could not account for any such large change in particle size.

A third possibility which presents itself is the possibility that a charge differential exists between the polystyrene sphere and its surroundings. If the spherical particle is charged negatively with respect to its surroundings, the potential gradient around the particle would constitute a negative electron lens acting on those electrons passing through and near the particle. Such a lens action would result in the size of the image of the sphere being smaller than would be expected if the sphere were not charged.

Shadowing the specimen at a low angle with a conducting film of platinum-palladium provides an electrical contact which substantially eliminates any charge differential between the particle and its surroundings. This reduces or eliminates the lens effect. The resulting image would be larger.

If this is correct, vertical shadowing of the spheres in such a way that no electrical contact is made between the metal cap on the sphere and the metal deposited on the film should result in no decrease of the charge on the particles, and the image size should be that of unshadowed particles. Figure 3 shows three electron micrographs of the same field—Fig. 3a before shadowing, Fig. 3b after vertical shadowing, and Fig. 3c after oblique shadowing. A just measurable increase in the size of the spheres was found for the vertical shadowed balls. This could be accounted for in that the amount of metal deposited on the ball and film in vertical shadowing was much greater than for oblique shadowing. The obliquely shadowed spheres in Fig. 3c are larger than the unshadowed and vertically shadowed spheres.

It appears that this experiment substantiates the charge theory; and it has been observed, also, that there is a difference in the position of the Fresnel diffraction lines on the shadowed and unshadowed side of a polystyrene sphere for any given setting of the objective lens current. This can be interpreted to mean that the shadowed side and the unshadowed side of the sphere are not brought simultaneously into the best focus.

Close observation of fields having spheres lying partially within the shadows of other spheres (see Fig. 3c) reveals that the portion of the sphere within the shadow of another sphere, and hence not shadowed, has a smaller radius of curvature than the adjoining portion which is shadowed. The sharpness of this effect leads one to believe that whatever the ultimate basic cause of these observations it must be well defined and intimately related to the presence or absence of a continuous film of metal.

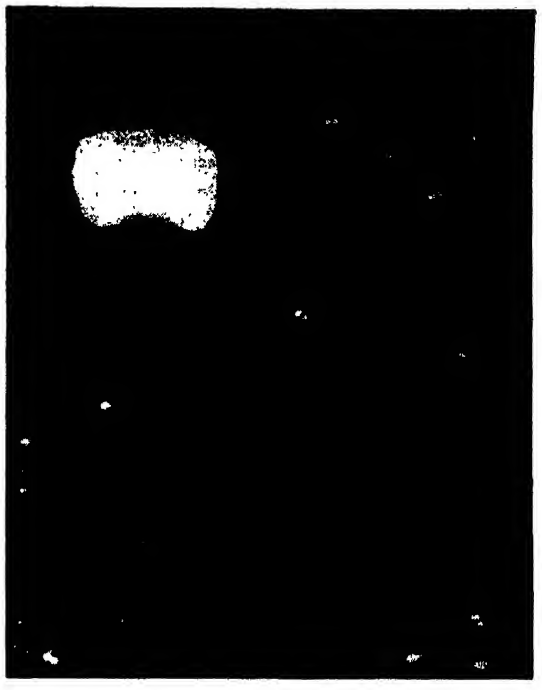


FIG. 4. *Streptomyces Griseus* Phage. The size differences between the shadowed and unshadowed phage are more than can be accounted for by the added platinum-palladium. 33,750×.

In spite of these findings it is still possible to use the spheres as a rapid check on calibration as well as for an internal standard of magnification calibration, if one first ascertains the image size to be expected for the condition under which the spheres are used.

The question arises whether objects other than polystyrene latex particles are also subject to this phenomenon. In Fig. 4 is shown one of the phages of *Streptomyces Griseus*. A few are within the shadow of the conidium; and therefore they are not shadowed. Others are outside the shadows and are themselves shadowed. The shadowed bodies are obviously much larger than the unshadowed. This is an indication that the same electron optical condition demonstrated for the polystyrene latex particles also affects the image size of these phage particles. If the present qualitative interpretation of these observations is correct, it is anticipated that this inherent error in apparent size of unshadowed objects in the electron microscope may be very generally encountered.

ACKNOWLEDGMENT

The authors wish to acknowledge the criticisms and suggestions made by Dr. W. W. Davis during the progress of this work.

The Current-Hysteresis of the Electrolytic Aluminum-Oxide Rectifier

A. J. Dekker and Helen M. A. Urquhart
*Physics Department, University of British Columbia, Vancouver, British Columbia, Canada**
(Received January 3, 1950)

A hysteresis in the rectification of the contact aluminum-aluminum oxide-electrolytic solution was found by Dekker and van Geel. In the present paper, measurements of the current-voltage characteristics between 1 and 2000 cycles are described. A tentative explanation of the hysteresis is given in terms of a pre-breakdown phenomenon.

I. INTRODUCTION

WHEN, in a solution of ammonium-borate, an aluminum plate is used as anode and an inert counterelectrode as cathode, the current, for constant applied voltage, decreases with time until only a small current remains flowing.[1] During the decrease of the current an oxide layer is formed on the aluminum. The final thickness of the layer is approximately proportional to the applied voltage (formation voltage) and amounts to about 0.1 micron per 100 volts.[2] During the formation process the current is carried almost completely by ions migrating under the influence of the high electric field in the layer[3] (10^6-10^7 volts per cm). The small "leak-current" observed after the formation process has ceased is probably caused by electrons flowing from the solution through the oxide into the aluminum.[2] Because of the high capacity per unit area of these layers, they find application in electrolytic condensers. The so-called "porous-layers" will not be discussed in this paper.[4]

It is well known that the system aluminum-aluminum oxide-electrolyte exhibits rectifying properties: the currents are very small for the aluminum positive (below the formation voltage, of course) with respect to the solution, while for opposite polarity they may be very large. It may be noted in passing that zirconium and tantalium show similar properties. The relation between current and voltage of the system is treated in this paper. In particular the behavior of the rectifier when a.c. voltages are applied will be discussed.

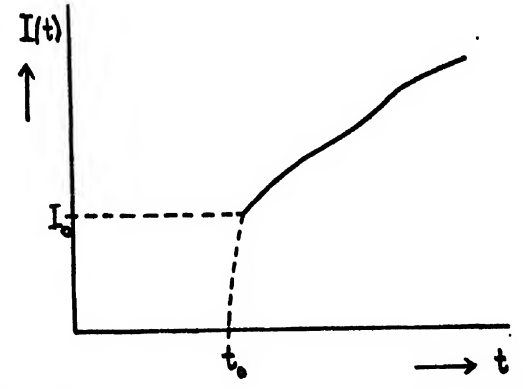


Fig. 1. The current I of an electrolytic rectifier as a function of time, for a d.c. voltage applied in the forward direction at a time $t=t_0$.

The d.c. current-voltage characteristics have been the subject of many previous investigations.[5] However, the origin of the observed time-effects discussed in Section II below, is not completely understood. Investigations of the a.c. characteristics, mainly at 50 cycles, were started by Dekker and van Geel.[6] These authors found a large hysteresis loop in the forward current (aluminum negative). In the present paper the a.c. characteristics for frequencies ranging from 1–2000 cycles per second will be discussed.

II. D.C. CURRENT-VOLTAGE CHARACTERISTICS

Since the d.c. current-voltage characteristics have a bearing on the behavior of the electrolytic rectifier for a.c. voltages, they will be briefly reviewed. It will be assumed that only voltages below the formation voltage are applied to the rectifier, so that no further oxidation takes place.

Suppose that at time $t=t_0$ a constant voltage V is applied in the forward direction (Al negative). The forward current generated by V is not constant but varies with time (Fig. 1). The current reaches a certain "instantaneous" value I_0 by the time a meter responds (approx. one sec.) and then increases slowly but erratically with time. Because no saturation value of the current is reached over a period of several hours, it is possible that during this time chemical changes occur in the layers that lower their resistance. This change of resistance will be referred to as the after effect. What happens between $t=t_0$ and the time when the meter responds will be discussed below (Section VI). The "instantaneous" value I_0 introduced above is determined not only by V but also by the "history" of the

* This research was carried out with the aid of a grant from the Defense Research Board at the University of British Columbia. The authors are much indebted to this institution for allowing them to publish this work.

[1] W. Ch. van Geel, Zeits. f. Physik 69, 765 (1931).

[2] W. Ch. van Geel and H. Emmens, Zeits. f. Physik 87, 220 (1933).

[3] E. J. W. Verwey, Physica 2, 1059 (1935).

[4] See W. Ch. van Geel and H. Emmens, Physica 1, 45 (1934); S. Anderson, J. App. Phys. 15, 477 (1944); A. J. Dekker and W. Ch. van Geel, Philips Res. Rep. 2, 313 (1947).

[5] See for instance W. Ch. van Geel, reference 1; A. Günther-schulze and H. Betz, Zeits. f. Elektrochemie 37, 727 (1931); R. Störmer, Zeits. f. Tech. Physik 16, 508 (1935).

[6] This research was carried out at the Philips Research Laboratories, Eindhoven, Holland.

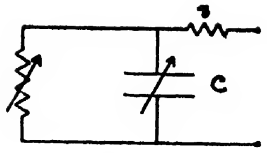


FIG. 2. Equivalent circuit of a rectifier.

rectifier. In general, voltages applied in the forward direction before $t=t_0$ tend to increase the "instantaneous" value, while voltages in the reverse direction tend to decrease it. This effect will be referred to as the hysteresis. For each forward voltage V a set of $I(t)$ curves may be obtained, each curve corresponding to a different "history" of the plate, that is, corresponding to events in the time between $t=-\infty$ and $t=t_0$.

Similar phenomena occur when d.c. voltages are applied in the reverse direction. When for $t<t_0$ a voltage in the forward direction was applied to the rectifier and at $t=t_0$ a voltage V is applied in the reverse direction, the current will reach an "instantaneous" value and then decrease, usually within a few seconds, to a constant value.

It will be clear from the preceding discussion that the d.c. characteristics can be given unambiguously only by specifying the conditions under which they were measured. The most reproducible results are obtained when before each measurement of the "instantaneous" current the rectifier is kept at the formation voltage (in the reverse direction) for a few minutes. The current in either direction is then approximately proportional to the square of the voltage. The rectification ratio, that is, the ratio of the forward and reverse current for the same absolute value of the applied voltage, may be as high as 10^4 for a plate oxidized at about 300 volts. For the measurements described in this paper, the formation voltage was applied for a few minutes before the photographs of the a.c. characteristics were taken in order to make the initial conditions for each photograph nearly the same.

III. THE CAPACITY OF THE RECTIFIER

In connection with the method for measuring the a.c. current-voltage characteristics (see Section IV) it is necessary first to point out some features of the equivalent circuit of the electrolytic rectifier.

It is well known that the simplest equivalent circuit for a *dry* rectifier that describes the electric properties of such a metal-insulator-semiconductor contact, consists of a non-linear resistance R shunted by a non-linear capacitance C, the two in series with a linear resistance r (Fig. 2). The capacitance C arises from the insulating layer (chemical barrier layer) and a physical barrier layer (Schottky layer) existing in the semiconductor next to the insulator. The effective thickness of the latter depends on the applied voltage and hence leads to a non-linear capacitance. The voltage dependence of C is most pronounced for zero bias voltage. Also, the voltage dependence of C decreases with increasing thickness of the chemical barrier layer.

To investigate to what extent the circuit of Fig. 2 can be applied to the *electrolytic* rectifier, the following experiments were carried out: An aluminum plate which was 50 cm² in area, was oxidized to a formation voltage of 285 volts in a solution of ammonium-borate (sp. res. 490Ω cm). Tinfoil was used as counter electrode. The impedance of the system thus obtained was measured at zero bias voltage for frequencies between 60 and 15,000 c.p.s. The measurements were carried out by using a General Radio 716-*B* capacitance bridge and a General Radio 608-*A* oscillator. The output of the oscillator was so adjusted that the effective voltage across the rectifier was 0.5 volt. The results obtained could be accounted for over the whole frequency range by the circuit of Fig. 2 with the following numerical values: $C=1.65$ μf, $R=10^5\Omega$ and $r=23\Omega$.

The impedance of the rectifier was also measured with different bias voltages in the reverse direction. No changes in the capacitance could be detected. Similar measurements with bias voltages in the forward direction are practically impossible as a consequence of the after effect mentioned in the previous section. There is no reason to believe, however, that the capacitance

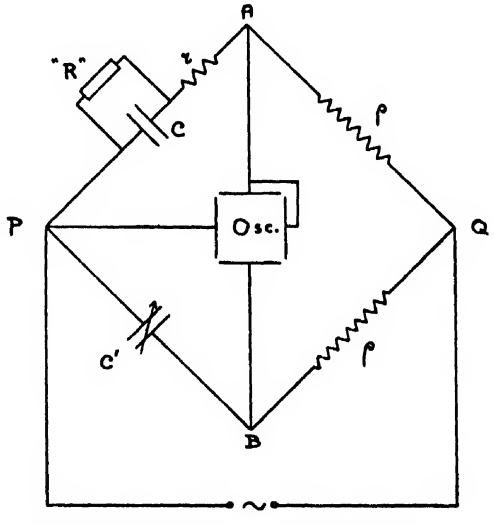


FIG. 3. Circuit for measuring the a.c. current-voltage characteristics.

would be different for voltages in the forward direction and therefore the capacitance will be treated as a constant in the following sections.

IV. METHOD FOR MEASURING THE A.C. CHARACTERISTICS

The a.c. current-voltage characteristics of *dry* rectifiers are usually obtained by putting a known resistance ρ in series. The voltage across ρ gives the current through the system and may be applied to the vertical deflection plates of an oscillograph. Using the voltage across the rectifier as the time base in the horizontal direction, the current-voltage characteristic may be readily observed.

FIG 4. Current-voltage characteristics at 2 (below) and 3 (top) cycles Current vertical, voltage horizontal

FIG. 5. From the bottom · characteristics at 3, 15 and 30 cycles

As a consequence of the presence of the capacitance C in the equivalent circuit (Fig. 2), however, $C\omega$ may become comparable with $1/R$ at high frequencies and the characteristic will then be distorted, especially in the reverse direction. The frequencies where this occurs are sometimes very high, for instance for germanium point-rectifiers (microwave region), whereas in other cases this region may lie in considerably lower ranges. It may be noted that Lehovec[7] has pointed out that this distortion in the reverse direction may be used to determine the capacitance of a rectifier.

For the *electrolytic* rectifier considered here, distortion due to the capacitance becomes noticeable at frequencies as low as 10 c p.s , as may easily be verified from the numerical values of C and R given in the preceding section. In this section a method to compensate for this effect, suggested by Dekker and van Geel, will be discussed.

In branch PAQ of the bridge in Fig. 3, the rectifier is represented by a constant capacitance C, shunted by a rectifying element "R" the nature of which is the subject of the present investigation, the two in series with the resistance r of the electrolyte. It should be noted that, because of the observed hysteresis (Section VI), the element "R" cannot be represented by a non-linear resistance only Suppose then that an a.c. voltage $V_0e^{j\omega t}$ is applied between P and Q. The current $I(t)$ flowing through "R" is some periodic function of time, depending on the input voltage, and may be developed in a Fourier series

$$I(t)=\sum_n I_n e^{jn\omega t}.$$

The voltage across the known resistance ρ may be calculated from the following differential equation

$$\rho C(d/dt)\{V_0e^{j\omega t}-(1+r/\rho)V_{QA}\}+\rho I(t)=V_{QA}.$$

Assuming V_{QA} can be expressed in a Fourier series, the solution is

$$V_{QA}=\rho\sum_n I_n e^{jn\omega t}\cdot\frac{1}{1+jCn\omega(r+\rho)}+j\frac{\rho C\omega V_0e^{j\omega t}}{1+jC\omega(r+\rho)}. \quad (1)$$

[7] K. Lehovec. J. App. Phys. 20, 123 (1949).

It is evident that for $\omega\to 0$ the voltage across ρ is directly proportional to $I(t)$ and hence in that case gives the current through "R" directly. It is also clear from (1) that for $\omega\neq 0$ the capacitance C gives rise to an "extra" capacitive current through ρ, leading to the second term in (1). This effect was compensated by Dekker and van Geel by using the branch PBQ, consisting of a variable capacitance C' and a series resistance ρ (equal to the one in the upper branch) in parallel with PAQ. Assuming that no current flows between A and B, the following equation holds for V_{QB}

$$\rho C'(d/dt)(V_0e^{j\omega t}-V_{QB})=V_{QB}.$$

This leads to

$$V_{QB}=j\frac{\rho C'\omega V_0e^{j\omega t}}{1+jC'\omega\rho} \quad (2)$$

It is evident that for $C'=C$ and $\rho\gg r$, the second term of (1) becomes equal to V_{QB}. Hence, by using a capacitance C' equal to the capacitance of the rectifier, the voltage across AB becomes

$$V_{AB}=\rho\sum_n I_n e^{jn\omega t}\frac{1}{1+jCn\omega\rho} \text{ for } r\ll\rho. \quad (3)$$

This expression shows that V_{AB} gives a picture of $I(t)$ which is distorted in two ways: the phase of each component of the Fourier expansion is shifted and, besides, the higher terms are more suppressed than the first term. However, the hysteresis loop does not show any discontinuities or peaks and, hence, only a few terms of the expansion are of importance. Therefore, V_{AB} will be approximately proportional to $I(t)$ if the condition $C\omega\rho\ll 1$ is fulfilled. For a given rectifier this condition puts a limit on the frequency range for which the method can be used.

The adjustment of C' can be made simply by measuring the capacitance C as indicated in Section III. A more convenient method of adjustment is the following: in the reverse direction $I(t)$ is practically zero for sufficiently low voltages. This means that C' has to be so

adjusted, that $V_{AB}=0$ in the reverse direction. Both methods gave the same value for the compensating capacity within the experimental error.

It should be remarked that complete compensation of the second term in (1) may be obtained by using between P and B in addition to C', a resistance equal to r in series. This becomes mainly important at very high frequencies where $C\omega\rho$ becomes equal or larger than unity. There, however, the method breaks down for another reason given above. The effect of not using a compensation resistance for r at high frequencies may be seen from (1): In the reverse direction, for $I(t)=0$, expression (1) yields

$$V_{QA}\simeq(1-r/\rho)V_0e^{i\omega t}, \quad \text{also} \quad V_{QB}\simeq V_0e^{i\omega t}.$$

Therefore, the voltage V_{AB} in the reverse direction, instead of being zero and leading to a horizontal line for the characteristic, will now give rise to a line with slope r/ρ. This was confirmed by the experiments (Fig. 6) described in the next section.

V. RESULTS

For investigations at frequencies below 60 cycles an oscillator of the Wien-bridge type, driving a push-pull amplifier with cathode-follower output, was built. Frequencies down to about one cycle/sec. with output of about 50 volts could be obtained with this apparatus.

Typical examples of characteristics at 2 and 3 cycles obtained with the circuit of Fig. 3 are given in Fig. 4. The rectifier consisted of an aluminum plate 50 cm² in area, oxidized in an ammonium-borate solution. The capacity of the rectifier was 0.80 μf. The resistances ρ were both 50 ohms. The solution in which the characteristics were measured had a specific resistance of 76 ohm-cm, leading to a series resistance r of about 0.5 ohm. Graphite counter-electrodes were used. For 3 cycles $C\omega\rho=8.10^{-4}$; hence (Section IV) the capacity has practically no influence on the shape of the loop obtained. The spot on the screen of the oscillograph follows the lower part of the loop for increasing forward voltage and returns along the upper part for decreasing forward voltage. This is in agreement with the results obtained by Dekker and van Geel at 50 cycles. In the reverse direction there is practically no current flowing and the characteristic becomes a straight horizontal line. The essential difference between Fig. 4 and the characteristic of a dry rectifier is of course the presence of the loop in the forward direction. These experiments show clearly that the loop is not due to a hysteresis in the capacity but arises from a time lag in the conduction mechanism. The characteristics are well reproducible if the formation voltage is applied for a short time before each photograph is taken.

At higher frequencies of, say 30 cycles, the following is observed: if no compensation capacity is used ($C'=0$ in Fig. 3), the characteristic in the reverse direction, instead of being a horizontal line, now "opens" up in a similar fashion as for a dry rectifier at high frequencies. This indicates that the effect is a consequence of the capacity C, and is confirmed by the fact that by using the circuit of Fig. 3 with $C'=C$, a straight line is again obtained in the reverse direction.

In Fig. 5 some results are shown for the plate at 3, 15 and 30 cycles. The effective voltage in each case was 34 volts. It will be noticed that the currents decrease with increasing frequency. This was observed also for other frequency ranges. The peak value of the current in the picture for 30 cycles in Fig. 5 corresponds to about 20 mA.

With increasing voltage, the general shape of the loop does not change. As a consequence of the non-linearity of the system, however, the peak value of the current increases more than linearly and the loop as a whole changes accordingly. The results of Dekker and van Geel gave a peak current approximately proportional to $V^{\frac{1}{2}}$. At higher voltages a loop in the reverse direction also becomes noticeable.

At high frequencies there is an advantage in using small plate area, since the capacity C and hence the distortion of the characteristic will be reduced in this way. There is also the advantage that the currents flowing remain small and hence the internal resistance of the oscillator used does not lower appreciably the voltage in the forward direction. That this happens may be seen from Figs. 4 and 5. There is, however, the disadvantage of increasing r and less sensitivity. As a consequence, compensation for r as well as amplification of the signal V_{AB} become necessary, unless the value of ρ is increased.

Some examples of a series of characteristics obtained for a small strip of oxidized aluminum are given in

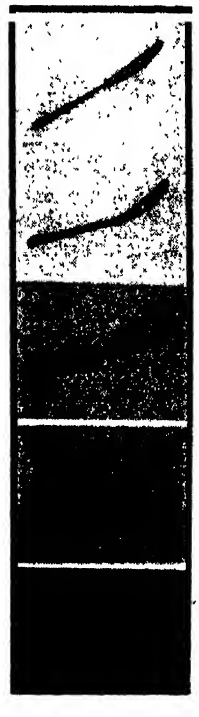


FIG. 6. Examples from a series of characteristics obtained with a small strip of oxydized aluminum. From the bottom: 20, 30, 150, 200, and 2000 cycles. No compensation for the series resistance was used in this case.

FIG. 7. Characteristics obtained for a strip of oxidized aluminum. From the bottom: 200, 300, and 400 cycles. In this case the series resistance of the electrolyte was compensated.

Fig. 6. The strip was oxidized to a formation voltage of 260 volts. Its capacity was 0.031 μf and the specific resistance of the solution was 76Ω cm. An effective voltage of 50 volts was used for all the photographs. Here again it is clear that the currents decrease with increasing frequency. Because only a capacity was used between P and B in Fig. 3, the effect of the series resistance r shows clearly: the characteristic in the reverse direction is no longer horizontal, but has a slope increasing with frequency. That this is indeed due to r may be seen from Fig. 7, where three pictures are shown obtained with a similar strip, but with a resistance to compensate for r shunting the capacity C'.

VI. DISCUSSION

The results obtained show definitely that the hysteresis loop is due to the conduction mechanism of the rectifying element "R" (Fig. 3). Therefore, any theory of the electrolytic rectifier should contain a mechanism that establishes a time lag between the current and the applied voltage. Logically, this time lag may either be caused by chemical changes in the layer which affect its conductivity, or, by processes of a physical nature that involve time, as, for instance, the building up of space charges in certain regions of the oxide.

In connection with the first possibility, experiments carried out by Störmer[8] are of interest. After preparing an oxide layer on an aluminum plate by anodic oxidation in an electrolytic solution, he applied a voltage in the forward direction for a certain length of time and used the plate thus prepared as a dry rectifier with aquadag as counter electrode. Störmer found that this treatment could improve the conductivity in the forward direction by factors up to a few hundred. He interprets his results by assuming that during the time that the forward voltage is applied to the plate in the solution, oxygen ions of the oxide are bound by hydrogen ions from the electrolyte. This leads to a semiconducting material with excess aluminum. This idea is compatible with experiments of Hartmann,[9] who found that aluminum oxide after certain heat treatments can become an excess semiconductor.

The chemical reduction, suggested by Störmer, seems the most reasonable explanation for the after effect referred to in Section II. For frequencies of the order of one cycle sec. or higher, however, it does not seem likely that changes of this type will occur and therefore the following "physical" interpretation of the hysteresis is tentatively put forward:

Suppose an aluminum plate is oxidized with a formation voltage of 300 volts. The thickness of the layer is then about 0.3 micron. For a voltage in the forward direction of, say, 30 volts, the field strength is about 10^6 volts per cm and is of the same order as the breakdown voltages for many ionic compounds. Electrons entering the oxide from the metal side (forward voltage) are accelerated by the field and may reach sufficient energy to eject electrons from energy levels below the conduction band. In view of the fact that the thickness of the layer is large compared with the mean free path (about 10^{-7} cm), the same reasoning may be applied to the secondary electrons and thus a multiplication process takes place. As a consequence of this process, positive holes are created, the number of which will increase with the distance from the metal.

In connection with the nature of these holes, it is important to note that the color of the luminescence, observed under certain circumstances during the anodic oxidation, depends on the kind of impurities contained in the aluminum. It seems reasonable, therefore, to assume that in the oxide layer localized electron levels between the highest filled band and the conduction band are present.[10] Electrons ejected from these localized levels will leave immobile holes behind and give rise to a space charge. As a consequence of the multiplication process, the positive space charge density increases with increasing distance from the metal. Also, the field at the metal-oxide interface will be larger than without space charge. Therefore, when a constant forward voltage is applied to the rectifier, the current increases with time until a stationary state is established in which the number of holes produced equals the number of electrons trapped by the holes.

Suppose then that an a.c. voltage is applied across the layer and that no space charge is present initially. For small voltages in the forward direction, the electrons will not gain sufficient energy from the field to produce secondaries and the field will be constant throughout the layer. When, however, the voltage reaches a certain critical value, a space charge will be built up. This explains why the forward current remains small for low voltages and then increases rapidly in the neighborhood of the peak value. During the subsequent decrease of the forward voltage, a certain time is involved in neutralizing the space charge so that a time lag between current and voltage is established.

[8] R. Störmer, Zeits. f. Tech. Physik **16**, 508 (1935).
[9] W. Hartmann, Zeits. f. Physik **102**, 709 (1936).
[10] See S. Anderson, reference 4.

On the basis of the pre-breakdown mechanism suggested here, it is also clear that the currents should decrease with increasing frequency. This is in agreement with the experimental results.

At very low frequencies, the after effect becomes important. According to the results obtained, there seems to be a continuous transition from the hysteresis to the after effect.

ACKNOWLEDGMENT

The authors wish to express their thanks to Dr. A. van der Ziel,* Physics Department of the University of British Columbia, Vancouver, British Columbia, Canada, for designing the oscillator for low frequencies and for many valuable and stimulating discussions.

* Now at the University of Minnesota, Minneapolis, Minnesota.

Scheme for Analyzing Capillary Measurements on Non-Newtonian Liquids

ORRIN H. CLARK AND M. L. DEUTSCH

Socony-Vacuum Laboratories, Research and Development Department, Paulsboro, New Jersey

(Received January 12, 1950)

A powerful new scheme is presented for analyzing capillary flow measurements on non-Newtonian liquids. The basic disability of the capillary method, namely its inability to assign to a flow measurement any particular shear rate, or even a known range of shear rates, is overcome. This is accomplished by developing simple relations connecting the directly measured quantities, pressure and flow rate, with viscosity and shear rate. Application to Bingham bodies is easy and an illustration is given. Application to measurement at high rates of shear is immediate in principle, but there are certain allied measuring difficulties. These are discussed briefly. Finally, a simple approximate formula applicable to cases of moderate deviation from Newtonian behavior is given.

I. INTRODUCTION

A POWERFUL new scheme is presented for analyzing the results of capillary flow measurements on non-Newtonian liquids. The classic capillary tube viscometer suffers from one intrinsic disability when applied to liquids whose viscosity varies with rate of shear. It can only measure flow corresponding to a complete range of shear rate, from zero to a maximum. Hitherto, there have been no satisfactory means of translating measurements over such a range into viscosity values at individual shear rates.[1] The present paper develops a simple method of obtaining the desired individual viscosity values from measurements of flow over a range of applied pressure.

II. DEVELOPMENT OF ANALYSIS OF CAPILLARY FLOW MEASUREMENTS

The data obtainable from capillary flow measurements consist of values of volume flow rate corresponding to applied pressures, together with the geometric constants of the capillary tube. In the case of non-Newtonian liquids such data do not immediately reveal any information about viscosity, nor even about the range of shear rate which was applied. To obtain such information explicitly, it is necessary to develop a relation between the two measured variables, flow rate and pressure, on the one hand, and the desired variables, viscosity and shear rate, on the other.

The following symbols will be used: Q=flow rate, volume per unit time; P=applied pressure difference; l=capillary length; a=capillary radius; r=radius of any annulus within the capillary ($0\leq r\leq a$); v=velocity of flow at any annular distance r; S=rate of shear ($S=-dv/dr$); η=viscosity.

The connection between viscosity and shear rate is immediately obtained from the differential equation appropriate to capillary flow, in which viscosity may vary with radius:

$$-Pr/l=d/dr[\eta r(dv/dr)]. \tag{1}$$

Integrating, noting that considerations of finiteness require the integration constant to be zero, and replacing dv/dr by $-S$:

$$\eta=(Pr/2lS). \tag{2}$$

Specifically, at $r=a$ and for pressure P:

$$\eta(Pa)=[Pa]/[2lS(Pa)]. \tag{3}$$

Equation (3) gives the desired connection. However, the function, $S(Pa)$, is unknown in the case of non-Newtonian liquids. $S(Pa)$ is now obtained in terms of the (experimental) flow-pressure curve. The flow rate may be written:

$$Q=\int_0^a 2\pi rv\,dr. \tag{4}$$

An expression for v is obtained by noting that

$$v=\int_a^r \frac{dv}{dr}dr=-\int_a^r S\,dr. \tag{5}$$

This involves the assumption $v=0$ for $r=a$, i.e., there is no slippage at the wall.

[1] For example, see H. Green, *Industrial Rheology and Rheological Structures* (John Wiley and Sons, Inc., New York), Chapters 1–3.

It will be helpful to introduce two new variables, $\rho = Pr$, and $\rho_1 = Pa$. In a single experiment ρ_1 is constant but is variable when a range of pressure is considered. Now (4) becomes

$$Q = \frac{2\pi}{P^2}\int_0^{\rho_1} \rho v d\rho, \tag{6}$$

and (5) becomes

$$v = \frac{1}{P}\int_{\rho}^{\rho_1} S d\rho. \tag{7}$$

The last integral separates into two parts, one of which has constant limits.

$$v = \frac{1}{P}\int_0^{\rho_1} S d\rho - \frac{1}{P}\int_0^{\rho} S d\rho, \tag{8}$$

$$= (say) \frac{1}{P}[A(\rho_1) - A(\rho)]. \tag{8a}$$

Introducing (8a) in (6):

$$Q = \frac{2\pi}{P^3}\left[\frac{\rho_1^2}{2} A(\rho_1) - \int_0^{\rho_1} \rho A(\rho) d\rho\right], \tag{9}$$

since $A(\rho_1)$ does not depend upon ρ. Differentiating (9) with respect to ρ_1, after placing $P = \rho_1/a$, we find

$$\frac{dQ}{d\rho_1} = -\frac{3}{\rho_1} Q + \frac{\pi a^3}{\rho_1} \frac{dA(\rho_1)}{d\rho_1}. \tag{10}$$

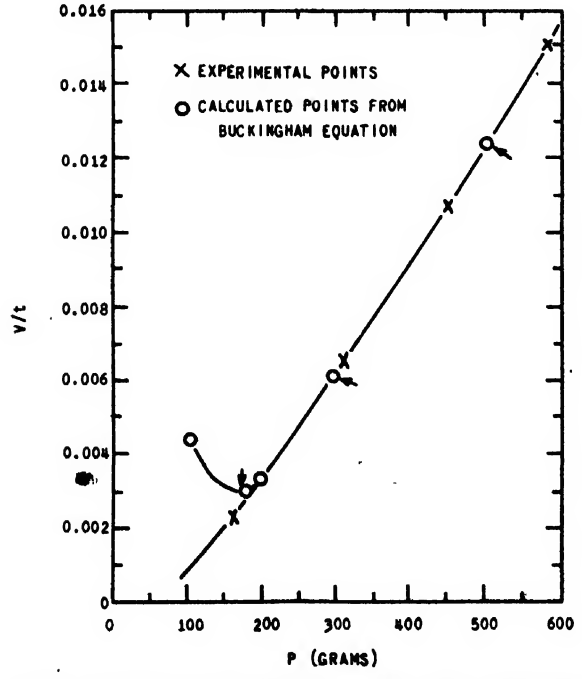


FIG. 1. Capillary flow data on a Bingham body. Reproduced by permission from *Industrial Rheology and Rheological Structures* by H. Green (John Wiley and Sons, Inc., New York, 1949).

Now noting the definition of $A(\rho_1)$,

$$dA(\rho_1)/d\rho_1 = S(\rho_1), \tag{11}$$

solving (10) for $dA(\rho_1)/d\rho_1$, and restoring the original variables we obtain

$$S(Pa) = 1/\pi a^3[3Q + P(dQ/dP)]. \tag{12}$$

Equation (12) specifies for any pressure P the shear rate (at $r = a$) in terms of directly measured quantities. An expression for the viscosity corresponding to this shear rate was given in Eq. (3). A complete graph of viscosity against shear rate may be obtained merely by plotting flow against pressure, obtaining the slopes dQ/dP at a number of pressures, and using Eqs. (3) and (12).

It will be noted that the shear rate given by (12) at each pressure is the maximum existing within the capillary tube, so that the range obtainable from the equations is the greatest possible, viscosity being obtained at the greatest shear rate in the liquid.

These results have great generality. It has been assumed there is no slippage at the wall. The only other limitation is implicit in Eq. (1), in which it is assumed that the applied pressure is opposed only by viscous forces, i.e., that there is no appreciable yield point or shear elasticity. This last restriction is easily removed. The yield stress bears a simple relation to the threshold pressure at which flow starts and is taken into account by a minor modification in the capillary flow equation. In the case of Bingham bodies the threshold pressure is not even required. This will be shown in the next section, which will serve to illustrate the method generally as well as the means of handling yield stress.

III. APPLICATION TO BINGHAM BODIES

Bingham bodies are characterized by a definite yield point (denoted by f), viscosity being constant after flow starts. In this case (1) must be modified to include the yield stress.

$$-Pr/l + f = d/dr[\eta r(dv/dr)]. \tag{13}$$

Integrating, and taking $r = a$:

$$f = Pa/2l - \eta S, \tag{14}$$

with S given by (12) as before. At the threshold pressure $S = 0$ and f is obtained at once. In the case of Bingham bodies any two points on the flow-pressure curve suffice to determine f and η; the constancy of η may be verified by additional points.

An example is given using data on pigment-vehicle mixtures published by Green.[2] These are reproduced in Fig. 1. Pressure values of 200, 300 and 500 are selected. Sufficient accuracy will result from using direct differences in the table to obtain slopes, and the 300 to 500 difference will be used for pressures of 300 and 500

[2] See reference 1. Reproduced by permission.

both. Then:

$$P=200,\quad dQ/dP=0.0000289,\quad S=159,$$
$$P=300,\quad dQ/dP=0.0000315,\quad S=287,$$
$$P=500,\quad dQ/dP=0.0000315,\quad S=547.$$

Using the values for $P=200$ and $P=300$ in (14), and changing grams to dynes gives:

$$f=422 \text{ dynes/cm}^2,\quad \eta=3.92 \text{ poises}.$$

The check using points $P=300$ and $P=500$ yields the values:

$$f=428 \text{ dynes/cm}^2,\quad \eta=3.86 \text{ poises}.$$

IV. HIGH RATES OF SHEAR

The extension of capillary flow measurements to high rates of shear encounters formidable measuring difficulties quite apart from the problem of variable shear rate which has been treated above. These difficulties arise from the necessity for high pressures and large amounts of internally developed heat, resulting in serious gradients in temperature and pressure within the capillary. Analysis of flow results including allowance for viscosity variation due to temperature and pressure effects is very difficult.

Since the starting point for the scheme just developed assumes the possession of true values of flow for given pressures, it apparently has no immediate application to these measuring difficulties. However, the development of an experimental technique for measurement at high shear rates has been pushed to an advanced stage at the Petroleum Refining Laboratory, Pennsylvania State College,[3] with the result that data are obtained which merit discussion from the viewpoint of analyzing the effect of variable shear rate.

Details of this experimental technique are not relevant to the discussion. It is sufficient to state that the method achieves the assignment of an "apparent viscosity" to each measured flow rate which is the viscosity of a Newtonian liquid producing the same flow rate at the same applied pressure. Approximate corrections for temperature and pressure effects have thus been incorporated in the apparent viscosity. For the purpose of comparative work the apparent viscosity is used directly to characterize the liquid.

If, however, a description of the fluid's non-Newtonian characteristic, i.e., its viscosity dependence upon shear rate, is desired, it is necessary to find means of assigning corresponding shear rates to the apparent viscosities resulting from capillary flow measurements. This is the very purpose for which the equations of Section II are designed. It is only necessary to translate the apparent viscosities into corresponding apparent flow rates by using Poiseuille's law. This will give the flow-pressure data needed and permit the construction of an explicit viscosity to shear rate relation.

[3] The authors are acquainted with this work through the courtesy of seeing the report to the American Petroleum Institute, P.R.L.-9-49, by Fenske, Klaus, and Dannenbrink.

TABLE I. Data for Fig. 1, experimental points.

P (g)	V/t (cc/sec.)	P	V/t
160.1	0.00230	100 calculated	0.00443
307.9	0.00652	174 calculated	0.00279
450.1	0.01084	200[a]	0.00320
588.2	0.01513	300[a]	0.00609
729.6	0.01963	500[a]	0.01239

Radius of capillary, 0.031365 cm
Length of capillary, 3.162 cm

[a] Points taken on experimental curve for calculating f and η.

It will be noted that no additional assumption is involved in applying Poiseuille's law to the apparent viscosities, for these are defined in terms of equivalence of flow rate.

A simplified procedure for assigning shear rates to the apparent viscosities is possible by introducing another Newtonian approximation. To the capillary flow in question there is assigned an effective shear rate defined as the shear rate effective in producing flow for Newtonian liquids. Thus:

$$\bar{S}=\frac{\int SdQ}{\int dQ}$$
$$=\frac{1}{Q}\int_0^a S2\pi rvdr$$
$$=\frac{1}{Q}\int_0^a \frac{Pr2\pi rP}{2l\eta 4l\eta}(a^2-r^2)dr$$
$$=(4/15)(Pa/l\eta). \qquad (15)$$

The validity of the approximation (15) will in general depend upon the magnitude of the deviation from Newtonian behavior. It should be noted, however, that in the case of the Petroleum Refining Laboratory work cited Eq. (15) constitutes an entirely satisfactory assignment of shear rate. It gives results agreeing within experimental error with those obtained by applying the method of Section II. However, one advantage does accrue from using the method of Section II; it makes possible the extension of the viscosity relation to substantially higher shear rates.

Interpretation of Fracture Markings

J. A. Kies, A. M. Sullivan, G. R. Irwin
Naval Research Laboratory, Washington, D. C.
(Received March 23, 1950)

This paper presents a study of the sequence of events by which cracks "grow." A careful investigation was made of the origin, development, and characteristic pattern of level-difference markings left by rapidly moving fractures in a wide variety of materials. Characteristic markings indicating discontinuous propagation are found in fractures of plastics, polycrystalline and single-crystal materials, coal, mica, photographic film from which the emulsion was stripped, etc. Thus, the continued occurrence of essentially discontinuous initiations which comprise fracturing becomes clear. These considerations are sufficient to explain a number of well known fracture markings and to clarify what is mean by brittleness and fracture velocity.

INTRODUCTION

THE detailed markings on fracture surfaces of metals, plastics, and ceramics show many common characteristics. A number of these are illustrated and reviewed in this paper for the purpose of showing what one may infer from these markings concerning the growth and propagation of fracturing.

LEVEL DIFFERENCES

De Freminville,[1] Preston,[2] and others have called attention to the importance of fracture level differences and crack directions as indicators of the stress pattern in glass. The mechanism by which level differences arise has been given relatively little attention. A comparative study of metals, transparent crystals, plastics, and glass, etc., sheds new light on this mechanism.

A disconnected initiation beyond the crack head is well illustrated by Fig. 1. This shows an aluminum foil under dead weight loading. The fracture is propagated by joining a sequence of holes originating at widely separated points. T. W. George and others at NRL have shown this to be the rule in the fracture of ductile foils.

For the purpose of studying fracture markings in thick specimens, lucite serves as a convenient model. Figure 2 shows the fracture surface of a Lucite plate, $\frac{1}{16}$-in. thick, broken in tension. One finds that the areas separated by the parabola-like lines are at different levels. Thus, these lines owe their visibility to the existence of sharp changes in level and may be termed "level-difference lines." More details of this are shown in Fig. 3 at higher magnification. Usually there is a barely resolvable pit or hole at the focus surrounded by a circular zone of high specular reflectivity. The streamers radiating from the focus and trailing off in the direction of the crack propagation are the boundaries of various sublevels within the parabola. It can be shown that the point of origin of the fracture of fractures making up the parabola was at or very near the the focus. Similar details can be found on finer and finer scale to the limit of resolution of the ordinary microscope. This leads to the inference that all level-difference lines or edges, especially those leading in the direction of fracture can be associated with separate initiations of fracture.

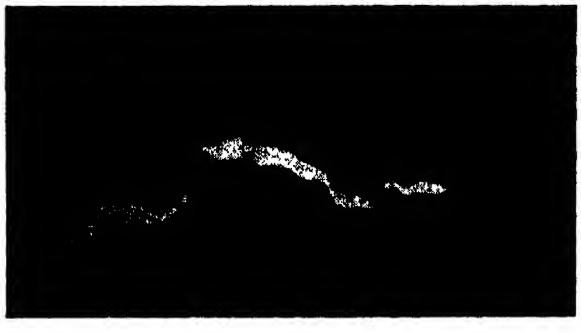


FIG. 1. (×15) Creeping fracture backlighted in 0.002-in. thick 2S aluminum foil (after T. W. George). The specimen had been under dead load for 30 days. The propagation consisted of the opening up of advance pinholes in severely strained regions followed by a joining process in which each subsidiary fracture spread backwards as well as forwards. Some such microfractures were not included in the joining process.

FIG. 2. Parabolic figures on the fracture surface of a Lucite plate (×75). Direction of crack propagation: Left to right. Each parabola outlines the area included in an individually initiated subsidiary fracture. The lines are visible becaues of differences in level.

[1] Ch. de Freminville, "Caractères des Vibrations Accompagnant le Choc Déduits de l'Examen des Cassures," Rev. de Metallurgie 4, 833 (1907). "Recherches sur la Fragilité: L'Eclatement," Rev. de Metallurgie 11, 971 (1914).

[2] F. W. Preston, "A study of the rupture of glass," J. Soc. Glass Tech. X, 234 (1926).

FIG. 3. Enlarged view of the foci of a group of parabolas from the preceding figure (×250). Satisfactory proof exists that fracture elements were initiated at the foci.

FIG. 4. (×7.5) Parabola head complete with focal pit and streamers on the fracture surface of a fine grained steel tensile bar. The vertex of the parabola points toward the spot where the main fracture first initiated. (Courtesy, Dr. W. Harris.)

FIG. 5. (×20) Enlarged view of the parabola shown in the preceding figure. At the focus is a small crater with sheared conical walls. The weakening influence presumably within the pit is not revealed.

The structural components of the parabola in lucite have individual analogs within the comparable portions of the cup-cone fractures in ductile steel. Figure 4 shows a fracture in a tensile bar, kindly supplied to us by Dr. W. J. Harris of the NRL Metallurgy Division. The main fracture originated at the lower right and at a later time another prominent weakness at another level gave rise to a similar process starting from the focus of the parabola occurring at the upper left of center. Note that the parabola in the steel, Fig. 5, at higher magnification has a volcano-like focal crater indicating shearing at the focus. This is better illustrated in Fig. 6. Surrounding the shear is a bright granular area with streamers analogous to the pattern of Fig. 3 in Lucite. This area in the steel surrounding the focus also resembles the "mirror" in a glass fracture. Inasmuch as a granular appearance on a fracture in steel has been associated with high speeds of fracturing and low energy absorption, one may conclude that the toughest part of this fracture element was at the focus and that, once some shearing was completed, the subsequent development occurred far more rapidly. The sharp transition back to a more ductile appearance may be explained by reasoning that, at least locally, the fast fracture entailed a sudden expenditure of stored elastic energy and sudden stress relieval. This resulted in a slowing down or momentary hesitation of the crack front.

SCALLOPED LEADING EDGE

In order to establish the time sequence of events in the generation of parabola-like figures and conchoidal markings, microscopic observations were made during the progress of fracture in cellulose acetates and in a large crystal of KBr. In the case of cellulose acetate two photographs, Figs. 7 and 8, serve to show the positions of the main crack at the instants of halt or very slow propagation in response to a controlled relieval of the force exerted on the knife blade used to propagate the crack. The arrow points to the focal spot of a parabola-like figure and at the second position new parabola heads have been created within the outline of the original one.

The leading edge was deeply scalloped with a succession of finer and finer scallops superimposed on the principal ones. Each main scallop was the leading edge of a parabola-like fracture element within which could be found an individual point of initiation.

In the case of the fracture in KBr a razor blade was forced into the material by a steady motion of a screw. In this case the scalloped leading edge of the crack more or less following a cleavage plane was essentially the same in appearance as in cellulose acetate. The main scallops were observed to progress with a rapid succession of pulses and halts self-controlled by the specimen.

Conchoidal fractures are all produced in this way except that the fracture surface is not constrained to follow a natural cleavage. It is important to recognize that such discontinuous propagation generally involves many, not just one, transitions in mode from fast to slow. During slow stages more time is allowed for plastic flow, hence a conchoidal fracture or more generally any brittle but pulsating fracture may involve local ductile

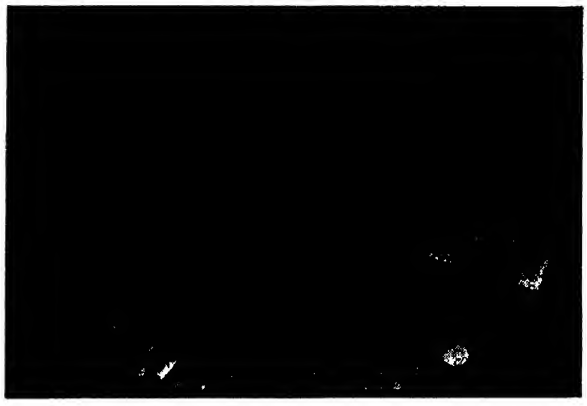


FIG. 6. (×7.5) The sector at the left was formed by radial crack propagation and corresponds to the parabolas of the preceding figures. A micro flaw is presumed to have been located at the focal pit lying within the volcano-like sheared structure. A small granular area surrounding the volcano suggests fast fracture at the expense of a local quick release of elastic energy. This was followed by slower more fibrous fracture at a somewhat lower mean stress. (Specimen supplied by Dr. W. Harris.)

FIG. 7. (×50) A fracture, traveling to the right, viewed within cellulose acetate. Note the well formed parabola head with focal point. A chisel is being forced into the material but the crack is momentarily halted. Note the lagging crack front at points where the level changes.

behavior. Klier has recently shown examples of this in ferrite.[3]

CHEVRON MARKINGS

The chevrons or herringbone in ship plate are easily modeled. Figure 9 shows an example of a tearing fracture in cellulose acetate. The concentration of gross fracture centers is greatest near the centerline. The fracture front obviously led at the center and lagged at the surface. However, chevrons need not point toward the origin of the running crack, provided the edges of the fracture can be made to lead. Figure 10 shows a plate of cellulose acetate in which side scratches promoted reverse herringbone.

[3] E. P. Klier, "A study of cleavage surfaces in ferrite," Pennsylvania State College, Mineral Industries Expt. Station, Final Report to ONR, October 15, 1948–1949.

Although the chevron pattern was widely considered to be indicative of brittle and therefore unsatisfactory behavior of the ship plates in failed merchant vessels, it should be understood that a considerable amount of plastic deformation was present. Tipper has pointed out that shearing was in evidence along the lines sometimes called herringbone.[4] More generally these are level-difference lines where the fracture lagged and hesitated. The final joining up of the levels by spreading towards one another involved the initiation of a new fracture front with an initially slow and therefore relatively ductile stage.

Our association here of high energy absorption with slowness of crack propagation rested partly upon the microscopic observations previously mentioned in connection with the fracture initiation foci within parabola heads in various materials. Such an association was thoroughly demonstrated on a grander scale by numerous dead loading and other more common experiments on various metal foils and plates of plastics in which central slots or scratches served to locate the beginning of fracture. In these experiments the work absorbed per unit length of crack was plotted against propagation speed and crack length. Invariably the initial or creeping stages of the crack growth required a rate of energy input several orders of magnitude higher than later, faster stages. This has been discussed by Irwin[5] and forms the main basis for his statement of the necessary and sufficient conditions for the onset of instability or the transition from ductile to brittle fracture.

GEOMETRICAL FIGURES

It is easy to show that simple geometrical figures on fracture surfaces are to be expected as a result of the interference between neighboring fracture elements.

FIG. 8. (×50) An advanced stage of the fracture, traveling to the right, shown in Fig. 7. The sides of the main parabola are greatly developed and newer, smaller parabolas have been created within the main one.

[4] C. F. Tipper, "The fracture of mild steel plate," Report No. R3, Admiralty Ship Welding Committee, London, 1948.

[5] G. R. Irwin, "Fracture dynamics," "Fracturing of metals," Am. Soc. Metals, 147 (1948).

FIG. 9. (×50) Herringbone or chevron fracture produced in a cellulose acetate plate. The fracture propagation was left to right. The heads of parabolas are clearly shown as concentrated at the center of the plate thus showing centrally located but discontinuous initation and propagation to be the reasons for such markings. The scalloped crack front was generally normal to the sides of the parabolas. Note that some figures become rapidly divergent and have hyperbolic outlines. This is due to local speed effects.

FIG. 10. (×7.5) Reversed chevrons or herringbone fracture in a cellulose acetate plate. Side scratches promoted initiation near the plate surface rather than in the center. As the fracture proceeded from left to right the fracture led at the surface and lagged in the middle contrary to normal chevrons in ship plate. Parabola heads are present but too small to be seen at this magnification.

For example, a rectilinearly propagating crack with a straight line front and a radially propagating crack with a circular front both in the same plane will meet on a conic section. If the velocities are equal, the figure is a parabola. If the circular- or disk-shaped crack is the slower, an ellipse will result. An example of this is shown in Fig. 11. A slight level difference for the two fractures makes the figure easily visible. In any case a relatively slow element will be surrounded and stopped by its faster neighbors. The ensiform tips in Fig. 11 thus represent extinctions of fracture elements. These are also junction points in river systems to be mentioned later. A hyperbola-like figure is to be expected as the boundary of a fast fracture element if it spreads more rapidly than its neighboring fracture elements. Such an outline is suggested at several places in Fig. 9.

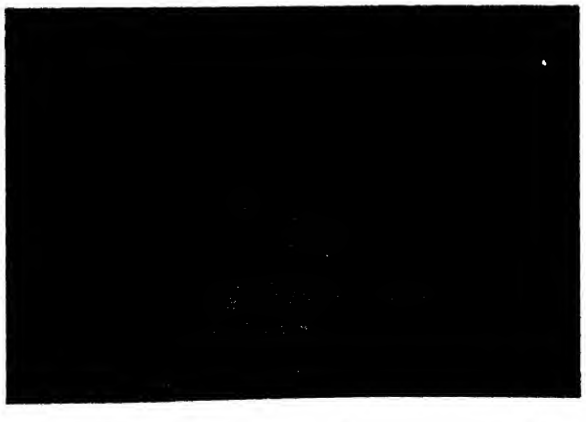


FIG. 11. (×50) Fracture surface in cellulose acetate created by a crack traveling right to left at 0.01 in./sec. Note the elliptical figure swept out by a fracture element which was engulfed and cut off by surrounding elements which ran ahead. Each ensiform tip is the trailing end of a fracture element which was crowded out by faster adjoining elements. Parabola heads are too small to show at this magnification.

River systems result from the consolidation of level differences as the faster elements shoot ahead and stop the slower ones. Many such examples have been shown without this explanation in publications by Zapffe and others.[6] A large scale example of river systems is shown in Fig. 12. This is a face spall in steel armor. The rod-shaped pieces lying within the overlap of neighboring fracture elements are often thrown off as fragments designated (in French) by De Freminville as "stries". A complete understanding of the origin of these shards and the separate fracture elements creating them cannot be gained by the observation of fractures in glass and hardened steel alone since in these materials the scale of events is too fine or is obscured by crystallinity. In plastics, however, each tributary of a river system can be seen to be a level difference between fracture elements originating from separate initiation points or operative flaws. In general, only one limb of the parabola-like figure associated with each initiation persists with enough level difference to be noted as a tributary in the system. The flaw-like nature of these initiations is demonstrated by Fig. 13 in which the boundaries of poorly molded particles of Lucite provide many such points for fracture initiation. Such examples lead to the inference that operative flaws play a promi-

FIG. 12. (×.5) Face spall in 4-in. thick Class *A* armor. The fracture initiated at the top of the picture. Note the banded structure with repeating river systems. Each tributary marks a difference in level between adjoining components of the fracture. Tributaries converge because some fracture elements were too slow to compete. Each band edge is a position of the crack front at an instant of hesitation in the propagation.

[6] C. A. Zapffe and F. K. Landegraff, "Fractographic examination of tungsten," Am. Soc. Metals 41, 396 (1949).

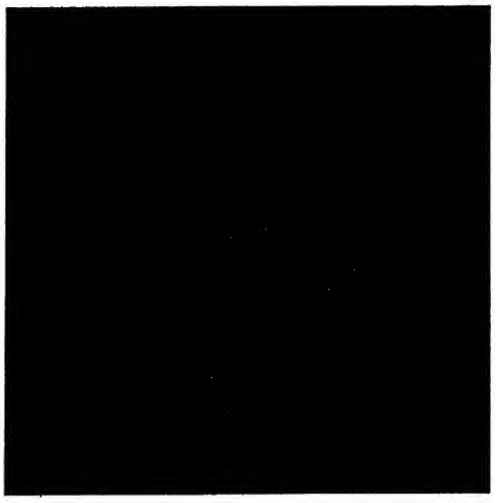


FIG. 13. (×250) "Grain boundary" weakness in imperfectly molded Lucite.

FIG. 14. (×50) Multiplicity of fracture initations contributing to a high speed shattering fracture in Lucite. The star-like components have spread more or less concurrently so that a system of honeycomb boundaries separates the components lying at slightly different levels. There is no evidence of a main crack front here. The presence of such a system is enough to insure low ductility and low dw/dx.

nent role in the propagation of fractures in all materials showing gross fracture patterns of the kinds described here.

EXPLOSIVE FRACTURE

The conditions of propagation as modified by high overstress can be indicated by the system of star-like fracture components shown in Fig. 14. Each one resembles the star in a tensile fracture of steel. Their independence of the main fracture results in their circular form. The regions between the initiations were evidently encroached on by many spreading fractures and thus little opportunity for plastic action was permitted. It is conceivable that operative flaws, e.g., lattice imperfections or dislocations, may be created anywhere in any solid material provided that there is sufficient time and an appropriate set of stresses. It follows that any material, however tough in static tests, can be made to fracture in a brittle manner provided that many flaws become operative simultaneously. This condition is readily achieved in suitable explosion tests of steel plate and hollow vessels. In such cases the over-all or apparent fracturing consists of many widely separated fractures joining together in more or less orderly sequence.

CONCLUSIONS

In conclusion, it may be stated that:

(1) The propagation of fracture is, in carefully observed cases, essentially and generally discontinuous and consists of the joining up of multitudes of separately initiated components of fracture.

(2) The more brittle the substance, the more striking are the effects of speed variations in creating markings usually called clam shell or conchoidal. Speed differences between components are responsible for the flowing together of markings resembling river systems and for the creation of geometrical patterns visible as parabolas, ellipses, and hyperbolas.

(3) Chevron or herringbone markings on ship plate may be explained on the basis of centrally concentrated gross initiation coupled with a centrally leading but scalloped crack front. Some initiations may be far in advance of the main crack and backwards propagation must occur locally.

(4) The transition from slow to moderately fast but not explosive fracture in hardened steel is accompanied by finer and finer scaled elements initiating closer and closer to the leading edge of the main crack as the local velocity of propagation increases. Fewer and fewer main elements persist with individually smoother faces.

(5) The over-all velocity of propagation of a crack is governed by the rate and distance at which advance initiations can occur and the rapidity with which subsidiary multiple fractures can be joined up. Many component joinings may occur simultaneously. This may be described as explosive fracture.

(6) In cases of high speed fractures traveling in high stress fields, the great number of fracture elements operating simultaneously tends to produce a condition in which the material in the path of the oncoming crack is permeated by multitudes of local stress raisers and flow restricters. Material surrounded by initiating fractures is subjected to a condition similar to that in a sharply notched tensile impact bar. Low ductility is a consequence.

The authors gratefully acknowledge the help and cooperation received from the following: Dr. W. Harris of the NRL Metallurgy Division, who supplied us with the tensile fractures in steel. Mr. C. Hauver, Head of the Photographic Section, Mechanics Division, NRL, who prepared some of the photographs. Messrs. H. L. Smith and W. J. Ferguson of the Mechanics Division, NRL, who performed many necessary fracture experiments contributing to this report. Dr. W. H. Sanders, Consultant, Mechanics Division, NRL, who assisted by editorial criticism and advice.

Letters to the Editor

An Absolute Measurement of Copper-Copper Interfacial Free Energy

G. W. Sears
General Electric Research Laboratory, Schenectady, New York
March 31, 1950

IT has been proposed by Smith[1] that the dihedral angles formed by three intersecting grain and phase boundaries are determined by a simple mechanical balancing of the interfacial tensions. He presented a mass of evidence to substantiate his assumption. On this basis he measured the ratio of copper(solid)-copper(solid) to copper(solid)-lead(liquid) interfacial free energies to be 0.59 from measurements of the dihedral angles formed at the junction of a pair of copper grains with liquid lead.

Following a suggestion of Fisher[2] the equilibrium angles at the edges of liquid lead drops equilibrated on a copper single crystal at high temperature were studied in an attempt to obtain an absolute value for the copper(solid)-lead(liquid) interfacial free energy and in conjunction with Smith's data an absolute value of the copper grain boundary free energy.

Outgassed cylinders of OFHC copper were polished on one end. The samples were placed polished side up in a hydrogen atmosphere and dusted lightly with lead powder. The specimens were heated for 8 hr. at 800°C.

After cooling, the samples were plated with 0.012 in. of copper. The specimens were cross-sectioned and polished normal to the lead-covered surface. The polished face was given a light dichromate etch and examined at 500× with a Bausch and Lomb metallograph.

After measuring approximately 500 dihedral angles at the edge of lead drops, it was established that they had distorted upon solidification while cooling to room temperature.

At the suggestion of W. Schockley it was assumed that shape of the drops remained undistorted upon solidification. On this basis the angles were calculated from measured lengths and thicknesses of the lenticular lead drops. The preliminary results on these measurements yield angles consistent with the random cross sectioning of lenticular drops normal to their median planes and having constant dihedral angles at their edges. Recent high temperature microscopic studies have independently confirmed Shockley's suggestion.

A preliminary value of 510 ergs/cm² has been obtained for the copper(solid)-lead(liquid) interfacial free energy. In addition a value of 720 ergs/cm² for the copper(solid)-lead(vapor) surface free energy was found. This differs from the value of 1370 ergs/cm for the copper(solid)-copper(vapor) surface free energy measured by Udin, Shaler, and Wulff.[3]

In conjunction with Smith's data a copper(solid)-copper(solid) interfacial free energy of 860 ergs/cm² has been determined. This represents the first absolute measurement of a solid-solid interfacial free energy to be reported in the literature.

[1] C. S. Smith, Met. Tech., T.P. 2387 (1948).
[2] J. C. Fisher, Met. Tech., T.N. 1 (1948).
[3] Udin, Shaler, and Wulff, J. Metals 1, 762 (1949).

Limiting Negative Pressure of Water

Lyman J. Briggs
National Bureau of Standards, Washington, D. C.
February 20, 1950

AT the 1949 meeting of the National Academy of Sciences in Washington I presented the results of some measurements, by a new method, of the limiting negative pressure of water. Improvements in technique have since led to values higher than those provisionally reported.[1] In particular, the investigation of the effect of temperature has disclosed a remarkable 10-fold change in the limiting negative pressure of water between 0° and 6°C.

The liquid under examination by this method is held in a capillary tube *open at both ends*. About 1 cm from each end the tube is bent back upon itself through an angle of 140° to form a Z. After thorough cleaning the tube is cemented (with thinned Duco) symmetrically to the web of a spinner of I-beam cross section. The spinner is mounted at its center on the upper end of the vertical shaft of a high speed three-phase induction motor, driven by a variable-speed motor-generator, equipped with an electric tachometer.

The motor and spinner are enclosed in a cylinder 25 cm in diameter, fitted with soft rubber gaskets at both ends and resting on the base supporting the motor. A tempered glass plate closes the top. When operating, the internal air pressure is reduced to 3 or 4 cm Hg.

The capillary tube spins in its Z plane, which is horizontal and is so mounted that its *center intersects the projected spin axis*. Consequently, the liquid column at this point is visible even at high angular speeds. The breaking of the column is signalized by a marked change in the refraction of the tube.

When the liquid under examination is an electrolyte, the rupture of the column can also be conveniently detected in the following way. A strip of filter paper 3 cm wide is cemented by its edges to the inside of the cylinder at the height of the spinner. Over this a continuous strip of open-mesh metal gauze 2.5 cm wide is cemented by its edges to the filter paper. The gauze and cylinder are connected in series with a 135-v B battery and a 25-v voltmeter. Normally the voltmeter shows no deflection. But the instant the column breaks and the electrolyte is thrown against the filter paper, a strong kick on the voltmeter is observed. In working with pure water, the filter paper is first activated by moistening with a solution of NaCl and then dried.

In the centrifugal field, one-half of the liquid column is pulling against the other. If one meniscus is nearer the spin axis than the other when the motor starts, it immediately retreats and the other advances until equilibrium is restored. The maximum stress is at the center. If L is the measured distance between the menisci in the bent legs of the tube, the effective length of the column is $r = L/2$.

The force exerted on any element of liquid in the column of length dx is $a\rho\omega^2 x\,dx$, where a is the cross section in cm², ρ is the density of the liquid, and ω is the angular velocity in radians per second. Integrating, from 0 to r, $F = \frac{1}{2}\rho\omega^2 r^2$ dynes/cm². The nega-

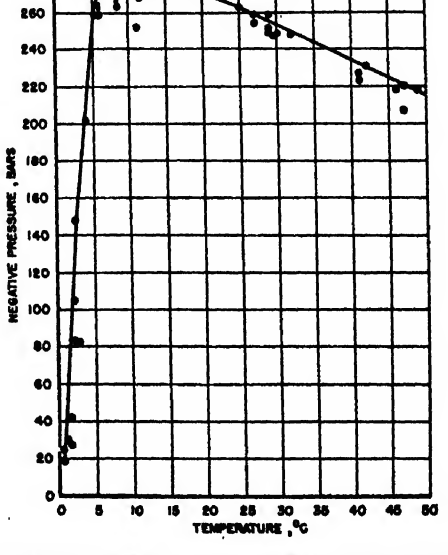


FIG. 1. The limiting negative pressure of water between 0° and 50°C.

tive pressure is given in bars (1 bar = 10^6 dynes/cm^2 = 0.987 standard atmosphere).

Scrupulous cleanliness is necessary in the measurement of negative pressure. The open-ended tube possesses a great advantage in being easy to clean and fill. To aid in these operations, the open ends of the tube project about 3 mm beyond the side of the supporting spinner.

The results of the measurements are presented in graphical form in Fig. 1. Points with arrows mean that the water column was subjected to the indicated negative pressure without rupture. It will be noted that the negative pressure is greatest (277 bars) around 10°C and decreases as the temperature is increased to 50°C. This decrease was expected, for at the critical temperature the negative pressure must be zero. The temperature coefficient from 25° to 50°C is about −2 bar/°C.

Between 5° and 0°C, the limiting negative pressure of water undergoes an enormous change, dropping to less than 10 percent of the maximum value. It presents another anomaly in the behavior of water in this interesting region. It cannot be accounted for by the change in density alone, for the density of water at 0° is only 0.00013 less than at 4°C; whereas a decrease in density of 0.009 occurs in raising the temperature of water from 25° to 50°C, which is accompanied by a change of only 20 percent in the limiting negative pressure.

The negative pressure measurements given in Fig. 1 are believed to be correct within three percent (one percent for ω and 0.5 percent for L). The temperatures corresponding to high values of the negative pressure are somewhat uncertain. The work done by the spinner on the residual gas in the chamber progressively raised its temperature and it is not known how closely the glass-enclosed liquid at the center of the column follows the temperature of the gas. Further measurements by a modified method are in progress.

The distilled water used in these measurements cannot be said to be air-free. The container from which the capillary tube was to be filled was completely immersed in distilled water and boiled vigorously. Inasmuch as the solubility of air in water decreases as the temperature is increased, it is assumed that the residual air in the boiled water was in a state of true solution at the lower temperatures at which the experiments were carried out.

Finally it must be emphasized that with this method we cannot tell whether the rupture of the column originates on the wall of the capillary (loss of adhesion) or in the body of the liquid (loss of cohesion). Strictly speaking, the measurements reported should be considered as applying to boiled water in a Pyrex glass capillary 0.6–0.8 mm internal diameter.

A fuller account of this work with additional data will appear in the *Journal of Research of the National Bureau of Standards*.

[1] Science **109**, 440 (1949).

The Effective Permeability of Iron Dust Particles

LEONARD LEWIN
Standard Telecommunication Laboratories, Ltd., Enfield, Middlesex, England
March 30, 1950

IN a recent paper on attenuators,[1] Theodore Miller quoted a formula on loaded materials from my paper "The electrical constants of a material loaded with spherical particles."[2]

As he says that the radii of the particles do not enter into the formula, I think there must have been some misunderstanding of the symbols I have employed, since the formula certainly contains the particle radius, although it is not shown explicitly.

The formula for the permeability, μ_m of the mixture is

$$\mu_m = \mu_b\left(1+\frac{3f}{(\mu_p+2\mu_b)/(\mu_p-\mu_b)-f}\right),$$

where μ_b is the permeability of the binder, usually unity, μ_p is the effective permeability of the particles at the frequency under consideration, and f is the volume ratio of particles to total volume.

The trouble arises over μ_p, which is *not* the permeability of the material from which the particles are made. If we call this latter permeability μ_2, then $\mu_p = \mu_2$ only when the radius of the particle is so small that the field completely penetrates the article. The general relation is

$$\mu_p = \mu_2 F(\theta)$$

where

$$F(\theta) = \frac{2(\sin\theta - \theta\cos\theta)}{(\theta^2-1)\sin\theta + \theta\cos\theta} \cong 1 + \frac{\theta^2}{10} + \frac{9\theta^4}{700} \text{ for } \theta \text{ small},$$

and $\theta = ka(\mu_2 K_2)^{\frac{1}{2}}$, $k = 2\pi/\lambda$ (λ is the free-space wave-length in cm), a = particle radius in cm, K_2 = complex permittivity of the particle material. For a metal $K_2 = -j60\lambda\sigma$, where σ is the conductivity in mho/cm. If the particle radius is much greater than the skin depth this gives $\mu_p = -j(\lambda/\pi a)(\mu_2/K_2)^{\frac{1}{2}}$, which is much smaller than μ_2. The general agreement which Miller obtains between theory and experiment probably points to the particle size being small enough for the approximation $F(\theta) \simeq 1$ to be used in most cases. The discrepancies actually recorded correspond to a value of F less than 1, which is in agreement with the formula for F, since θ^2 is negative for the values of the constants used.

[1] Theodore Miller, J. App. Phys. **20**, 878 (1949).
[2] L. Lewin, J. Inst. Elec. Eng. **94**, 64 (1947).

Magnetically Controlled Wave-Guide Attenuators

THEODORE MILLER
Electronics and Nuclear Physics Department, Westinghouse Research Laboratories, Pittsburgh, Pennsylvania
April 7, 1950

I WISH to thank Mr. Lewin for bringing to my attention an erroneous impression which I unwittingly created in a reference to his article which appeared in my paper "Magnetically controlled wave-guide attenuators."[1] I did not mean to imply that Mr. Lewin did not include the radii of particles in his general derivation for the effective permeability of a dielectric loaded with spherical particle, but rather that my application of his formula did not involve the radii.

Mr. Lewin's formula for the permeability of a loaded dielectric μ_m is[1]

$$\mu_m = \mu_b\left[1+\frac{3f}{(\mu_p-2\mu_b)/(\mu_p-\mu_b)-f}\right],$$

where f is the volume ratio of particles to total volume, μ_d is the permeability of dielectric, and μ_p is the high frequency permeability of the particles.

In his paper, Mr. Lewin relates the permeability of iron particles with the wave-length through an equation involving the particle radii, the conductivity, and the low frequency permeability of the iron material. However, since there was considerable uncertainty as to what values to assign to the conductivity and low frequency permeability of the iron material, it was decided to use Kittel's[2] value for the high frequency permeability of a thin film (2.5μ thick) instead of applying Mr. Lewin's equation which in part reads

$$\mu_p = \mu_2 F(\theta),$$
$$\theta = ka(\mu_2 K_2)^{\frac{1}{2}},$$
$$K_2 = -j60\lambda\sigma$$

where μ_2 is the permeability of iron material, σ is the conductivity, a is the radii of iron particles, and $k = 2\pi/\lambda$.

[1] Theodore Miller, J. App. Phys. **20**, 878 (1949).
[2] Leonard Lewin, J. Inst. Elec. Eng. **94**, 64 (1947), Eq. (11).
[3] C. Kittel, Phys. Rev. **70**, 281 (1946).

A Note on the Error Coefficients of a Servo Mechanism

JOHN L. BOWER
Department of Electrical Engineering, Yale University, New Haven, Connecticut
March 24, 1950

IN a linear servo system, it is convenient to be able to relate the error to the various disturbances on the system, the primary one of which is the input, θ_i. For a system initially at rest, one can write the following expression for that part of the error due to θ_i:

$$\epsilon(t)=\int_0^\infty d\lambda w(\lambda)\theta_i(t-\lambda). \tag{1}$$

The weighting function, $w(\lambda)$, is the inverse transform of the familiar transfer function relating the error to the disturbance, θ_i. For all physical systems, it includes a delta-function of order zero and of unit amplitude at the origin. If we assume θ_i to be a function which is differentiable through order n for all λ up to some λ_d, then we can write Maclaurin's formula for $\theta_i(t-\lambda)$ as a function of λ:

$$\theta_i(t-\lambda)=\theta_i(t)-\lambda\theta_i'(t)+\tfrac{1}{2}\lambda^2\theta_i''(t)+\cdots\frac{(-\lambda)^{n-1}}{(n-1)!}\theta_i^{(n-1)}(t)+R_n,\quad 0<\lambda<\lambda_d. \tag{2}$$

Here, R_n will be given the form,

$$R_n=\frac{(-\lambda)^n}{n!}\theta_i^{(n)}(t-\Delta\lambda)\text{ where }0<\Delta<1\text{ and }0<\lambda<\lambda_d. \tag{3}$$

We use the form $\theta_i^{(r)}(t-\lambda)$ to designate the rth derivative of θ_i with respect to the bracketed quantity. By inserting the identity (2) into (1), we have immediately an expression for error in terms of the derivatives of the input, θ_i, evaluated at t:

$$\epsilon(t)=C_0\theta_i(t)+C_1\theta_i'(t)+\frac{C_2}{2}\theta''(t)+\cdots\frac{C_{n-1}}{(n-1)!}\theta_i^{(n-1)}(t)+\frac{K_R}{n!}+K_d \tag{4}$$

where

$$C_r=\int_0^{\lambda_d}d\lambda(-\lambda)^r w(\lambda); \tag{5}$$

$$K_R=\int_0^{\lambda_d}d\lambda(-\lambda)^n\theta_i^{(n)}(t-\Delta\lambda)\cdot w(\lambda);$$

$$K_d=\int_{\lambda_d}^\infty d\lambda w(\lambda)\theta_i(t-\lambda). \tag{6}$$

It is obvious that the above series is most useful only when the last two terms are small enough to be neglected in the computation of $\epsilon(t)$. If the first m derivatives of the input are known to exist over all time, if the mth derivative is bounded by $\pm M$, and if m is sufficiently large that the corresponding remainder is negligible, we can represent the error by use of the coefficients up through C_{m-1}. The conditions here make for a simple calculation of the remainder, since $\lambda_d=\infty$, $K_d=0$, and

$$|K_R|\leq M\int_0^\infty d\lambda\lambda^m|w(\lambda)|. \tag{7}$$

Such a case as this might arise, for example, where the mth derivative is a stationary time series. The matter of evaluating the coefficients from (5) can be handled simply if the transfer function $H(s)$ relating the error to the input is known, for by definition of $w(\lambda)$,

$$H(s)=\int_0^\infty d\lambda e^{-s\lambda}w(\lambda). \tag{8}$$

By differentiating r times with respect to s, and taking s to zero in the result, we have

$$\left.\frac{d^rH(s)}{ds^r}\right|_{s=0}=\int_0^\infty d\lambda(-\lambda)^r w(\lambda)=C_r,\text{ for }\lambda_d=\infty. \tag{9}$$

These steps are clearly valid for any stable system, since its transfer function, $H(s)$, must be analytic in the neighborhood of the origin. The expression (9) places the coefficients of the power series in the familiar form shown in textbooks.[1] The inverses of C_0, C_1, and $C_2/2$, respectively, are referred to as the position-, velocity-, and acceleration-error coefficients of the servo.

For the ordinary rational fractional forms of $H(s)$ with simple poles, an upper limit for (7) can be obtained easily. We note that

$$\int_0^\infty d\lambda\lambda^m|w(\lambda)|=\left|\frac{d^m}{ds^m}\int_0^\infty d\lambda e^{-s\lambda}|w(\lambda)|\right|_{s=0}. \tag{10}$$

An upper bound for this quantity can be obtained by making a partial-fraction expansion of $H(s)$, substituting magnitudes of residues for the residue coefficients and replacing the poles by their real parts, after which the differentiation in s is performed and s is taken to zero. The result, placed in (7), is

$$K_R\leq M\int_0^\infty d\lambda\lambda^m|w(\lambda)|\leq M m!\sum_{k=1}^N\frac{|a_k|}{(R\alpha_k)^{m+1}}, \tag{11}$$

where the α_k are the locations of the N poles of $H(s)$, whose real parts are indicated by $R\alpha_k$, and whose residues are a_k. The form of (11) shows the tendency to weight most heavily the poles nearest the origin when m is large. For the usual servo these poles will be close to the real axis, which circumstance will make this bound on K_R approximate the value $M|C_R|$.

The series (4) may be useful even though the derivatives of sufficient order for application of the above method do not exist over all time. If the nth derivative fails at some λ_d which is sufficiently large that K_d from (6) is known to be small, and if (7) and (9) are good approximations for K_R and C_r, respectively, we can proceed as in the earlier case.

It might be mentioned that the above relations are valid for servo systems whose weighting functions, $w(\lambda)$ vary with time, t, as a parameter, in such a way that the law of superposition applies for each value of t.

[1] See James, Nichols, and Phillips, *Theory of Servomechanisms* (McGraw-Hill Book Company, Inc., New York, 1947), p. 147.

Erratum: Radiofrequency Mass Spectrometer

[J. App. Phys. 21, 143 (1950)]
WILLARD H. BENNETT
National Bureau of Standards, Washington, D. C.

FIGURE 12 of this article was improperly provided with a mass number scale which should have been as shown below.

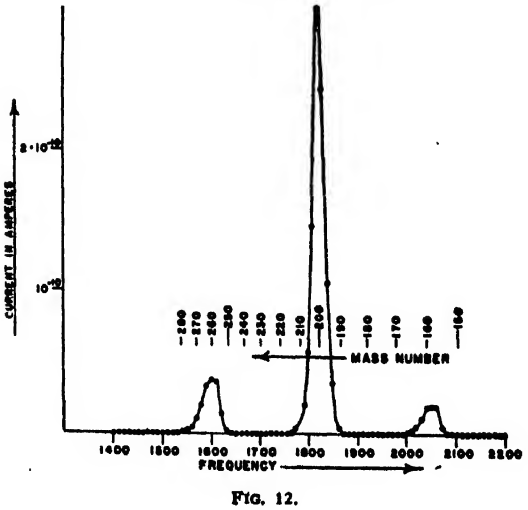


FIG. 12.

The Use of Critical Point Phenomena in Preparing Specimens for the Electron Microscope*

THOMAS F. ANDERSON
Johnson Research Foundation, University of Pennsylvania, Philadelphia, Pennsylvania
April 21, 1950

A SERIOUS handicap in the use of the electron microscope has been the fact that specimens must be dried before they can be examined in the high vacuum of the instrument. Delicate biological specimens are particularly susceptible to damage by the force of surface tension which flattens and sometimes disrupts them in the last stages of evaporation of the suspending medium. We report here a technique which eliminates the passage of a phase boundary through the specimen while it is being dried. After the specimen has been fixed, its water is replaced by alcohol, and the alcohol replaced in turn by each of a series of miscible liquids, the last in the series being one which has a conveniently low critical temperature. The temperature of the specimen in a bomb completely filled with the latter liquid is then raised above the critical point where the liquid changes imperceptibly into a gas. Then the gas is allowed to escape at the higher temperature leaving behind a dried specimen through which no phase boundary has passed. As a result, specimens like the hemolysed human red cells shown in the stereoscopic pair of electron micrographs of Fig. 1 retain their round shapes when dried in this way rather than being pressed flatly against the supporting Formvar membrane by the tension in the surface of an evaporating water droplet as they would have been in the conventional method of mounting specimens.

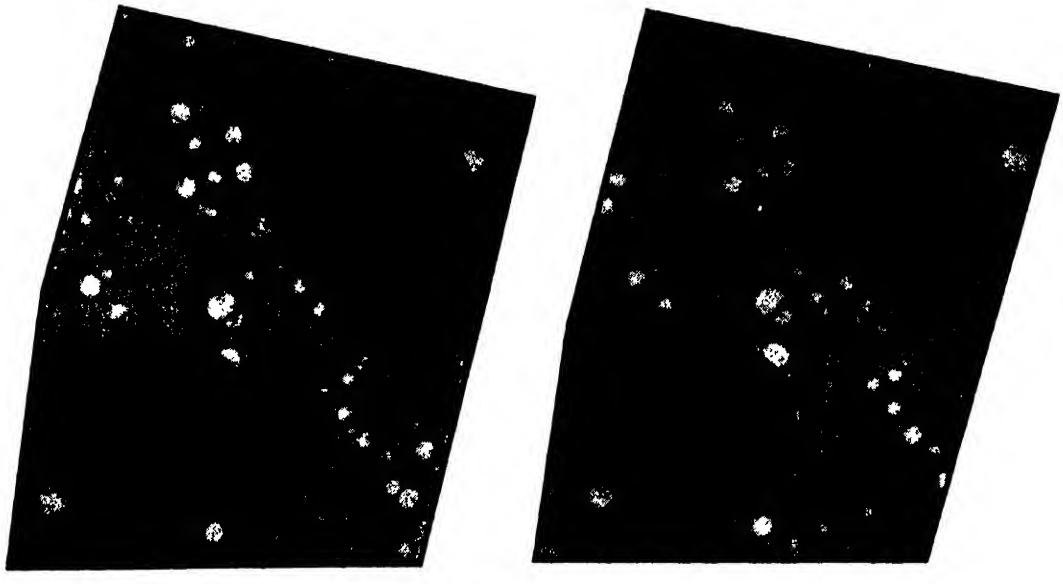


FIG. 1. Stereoscopic pair of micrographs of hemolysed red cells. Whole blood was diluted 1:10 in physiological saline, and the hemoglobin removed from the cells by the addition of an equal volume of water. A droplet of the hemolysed cells was then placed on a Formvar-coated specimen screen, fixed in osmic acid vapor, taken through alcohol to amyl acetate and then to liquid nitrous oxide in a pressure bomb at 25°. The temperature was then raised to 45° and the now gaseous nitrous oxide allowed to escape. The hollow "ghosts" of the cells may be seen adhering to the Formvar membrane.

* This work was supported by Contract N6 ori 168 T.O. II between the ONR and the Trustees of the University of Pennsylvania, and the Government is hereby granted royalty-free right of reproduction of this article.

International Congress on Electron Microscopy at Paris, September 14–22, 1950

A preliminary announcement has been received in this country concerning the above meeting. The following information has been abstracted from this announcement:

1. The Congress will include Sections on Electron Optics and Microscopes, Electron Diffraction, Metallurgical Applications, Chemical Applications, and Biological Applications as well as an Exhibition devoted to "The Microscope" and an Exhibition of the work of those attending.
2. The official language will be French but papers may be presented in English.
3. Titles and abstracts of papers should be submitted by June 1st to the General Secretary of the Congress, M. Locquin, Lab. de Cryptogamie, 12 rue de Buffon, Paris 5°, France.
4. The manuscripts of papers should be submitted by August 1st. The Congress reserves to itself the right to publish such manuscripts unless it has given a written release to the contrary.

Further details may be obtained from C. J. Burton, American Optical Company, Southbridge, Massachusetts.

Journal
of
Applied Physics

Volume 21, Number 8 August, 1950

Conductivity of Liquid Selenium—200°–500°C*

HERBERT W. HENKELS
Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania
(Received October 20, 1949)

The conductivity of liquid selenium has been measured in the temperature range 200°–500°C. The resistivity was expressed by $\log_{10}\rho = A + (B/T)$. Average values for different selenium lots and melts of A and B were -3.81 and 5850. The maximum deviations from the averages were 10 percent and 3.4 percent respectively. The resistivity was a function of temperature alone.

Various non-metal impurities Cl_2, I_2, P lowered the resistivities and produced different values of B in different temperature ranges. Mercury addition caused no change in either A or B although that metal greatly influences the resistance of solid hexagonal selenium. Melts doped with Cl_2 or Br_2, along with Hg exhibited behaviors different from those with single additions.

It was concluded that selenium is an ideal semiconductor in the range of measurement.

INTRODUCTION

STUDIES of the electrical properties of selenium are in progress at the University of Pennsylvania. Investigations on the subject have been summarized in Technical Report No. 9 entitled "Summary of Recent Work on the Conductivity of Selenium." It develops from a consideration of the data presented that a satisfactory explanation of conduction and the absence of measurable Hall effect[1] in solid selenium may require data on the conductivity of pure and doped, liquid and amorphous selenium. Existing data is scarce and unsatisfactory for reasons to be discussed later.

The study was considered profitable for several additional reasons. Satisfactory solid samples must at present be made according to an entirely empirical procedure. Studies of solid samples are greatly influenced by the crystalline condition of the sample. Therefore a detailed examination of the conditions of the selenium in all stages of sample preparation seemed highly desirable. Selenium is the simplest of the materials forming a stable amorphous phase. The data presented may prove useful in the further development of the theory of the liquid state. Commercial rectifiers are prepared from selenium doped with different halides. A continuous indication of the percentage impurity present in the molten selenium batch would be useful. The present report shows the possibilities of such an indication.

EXISTING DATA

Pelabon[2] reported a single measuring series with temperature variation from 390 to 690°C. The data was represented by:

$$\log\rho = 8.506 - 0.0095t(°C).$$

At 300° the resistance was noticed to decrease suddenly with a velocity that made measurements impossible. The origin of the selenium used in the experiments was not mentioned.

Borelius *et al.*[3] presented data on the resistance of solid and liquid selenium. Various selenium samples were sealed in glass tubes equipped with suitable platinum electrodes and the resistance was measured both as functions of periods of heating and of temperature. In the range measured, the specific resistance ρ followed the law, $\log\rho = A + (B/T)$, where T is the absolute temperature, A and B are constants. The results are given in Fig. 12.

* The following work has been done under contract NObs 34144 of the Bureau of Ships, United States Navy Department and previously reported as Technical Report No. 13 for that contract.

[1] H. W. Henkels, "Summary of Recent Work on Conductivity of Selenium," BuShips Contract NObs-34144, Technical Report No. 9, March 5, 1948.

[2] H. Pélabon, "Resistivity of selenium," Comptes Rendus 173, 295 (1921).

[3] Borelius, Pihlstrand, Anderson, and Gullberg, "Resistivity of selenium," Arkiv. Mat. Astron. Fysik 30A, No. 14 (1944).

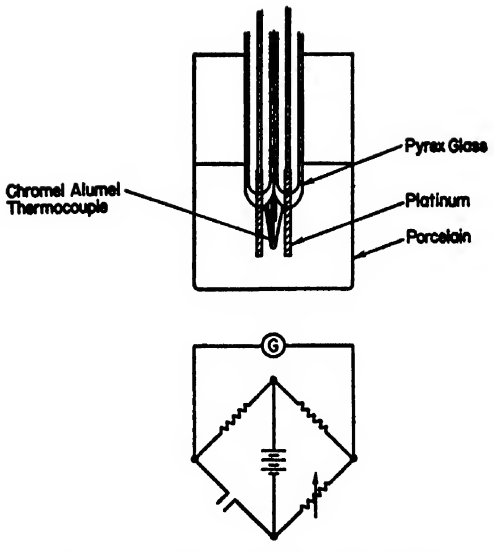


FIG. 1. Conductivity cell (liquid 200–500°C).

The values of A and B were observed to vary with the sample, period of heating and with depth of the electrodes in the melt. The heating periods were quite long (about 1800 hours). On the basis of the variation of resistance with depth of electrodes and differences observed in the conductivity of the cooling products of different strata of the melt, two liquid phases were postulated, a heavier one with low value of B and a lighter one with higher value of B. The properties of liquid selenium were adjudged similar to those of liquid sulfur.

NEW DATA

Source of Selenium

Selenium for the experiment was obtained from the Canadian Copper Refiners, Ltd. The analysis of the C.P. shot, performed by the laboratory of the company, was as follows:

Cu	Fe	Te	N.V.R.
0.1	0.8	0.2	5

parts per million. The selenium was purified in the following steps:†

1. Distillation at atmospheric pressure;
2. Oxidation in vapor state to SeO_2;
3. Resublimation of the SeO_2, followed by reduction in vapor state with NH_3 to elemental selenium;
4. Distillation of selenium in quartz apparatus, followed by condensation and shotting in distilled water.

TABLE I.

Sample	Curve	Temp.	A	B
lot 319 melt 1	$A, B, C,$	250–500°C	—	5650
lot 319 melt 2	D, E, F, G	250–500°C	—	5860
lot 319 melt 2	H, I	215–500°C	−3.99	5900–6010
lot 319 melt 3	N	250–500°C	−3.87	5930
lot 319 melt 4	W, X, Y	250–500°C	—	5930
lot 504-2 melt 1	AG	200–500°C	−3.50	5630
lot 504-2 melt 1	AH	200–500°C	−3.88	5882

† Information from Dr. C. W. Clark, Director of Research and Development, Canadian Copper Refiners, Limited.

Experimental Arrangements

Resistance was measured with a conductivity cell in an ordinary d.c. bridge circuit.

The conductivity cell is sketched in Fig. 1. Forty mil platinum wires were sealed in Pyrex tubing with 15 mm exposed. The glass tubes were then fastened together so that the electrodes were spaced 9 mm apart. A third glass tube with the end drawn to a small diameter and sealed, was fastened to the other tubes and served as the well for a thermocouple. This arrangement was placed in the center of a small porcelain beaker containing the sample which in turn was situated in the center of an electrically heated oven.

Decade resistance boxes were used as the arms of the bridge. These were calibrated and were found to have an error less than 1 percent. A Weston galvanometer with sensitivity of 0.25×10^{-6} amp./mm was used as null indicator.

For temperature measurements a chromel alumel thermocouple was introduced into the thermometer well. One wire was threaded into a piece of fine tubing to prevent shorting after the insulation had melted. The thermocouple hot junction was in contact with the thin drawn walls of the well. The e.m.f. was measured with a type K potentiometer whose smallest scale division was 5 microvolts. A Rubicon galvanometer with sensitivity

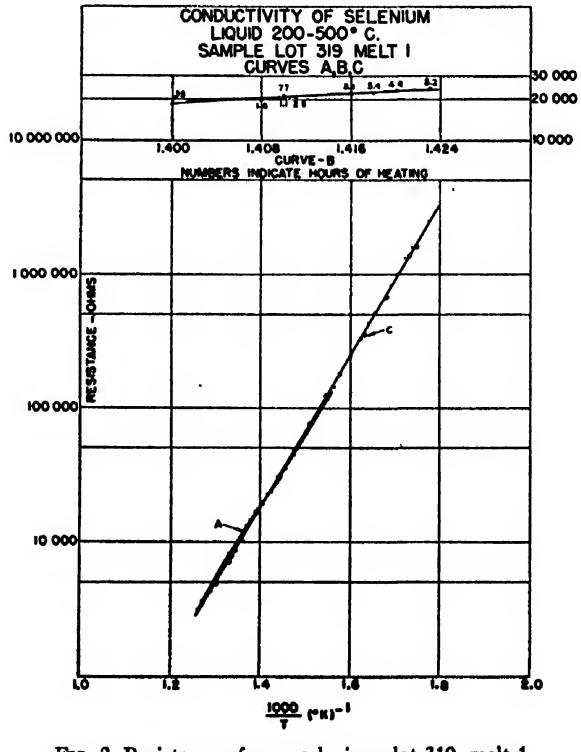


FIG. 2. Resistance of pure selenium, lot 319, melt 1.

of 6×10^{-10} amp./mm was used for null indicator. The standard e.m.f. was furnished by a new Eppley cell recently calibrated. Cold junction temperature was measured with a small mercury thermometer.

The conductivity cell constants were determined by calibration against an 0.01N KCl solution. The depth of the electrodes in the solution did not affect the value. The cell constants were the order of 1 ohm cm/ohm.

Consideration of Contact Resistance

Measurements on a cell with variable distance between the electrodes showed that contact resistance is of no importance.[3] This result agrees with the observations of Schweickert and Ioffe on the connection between

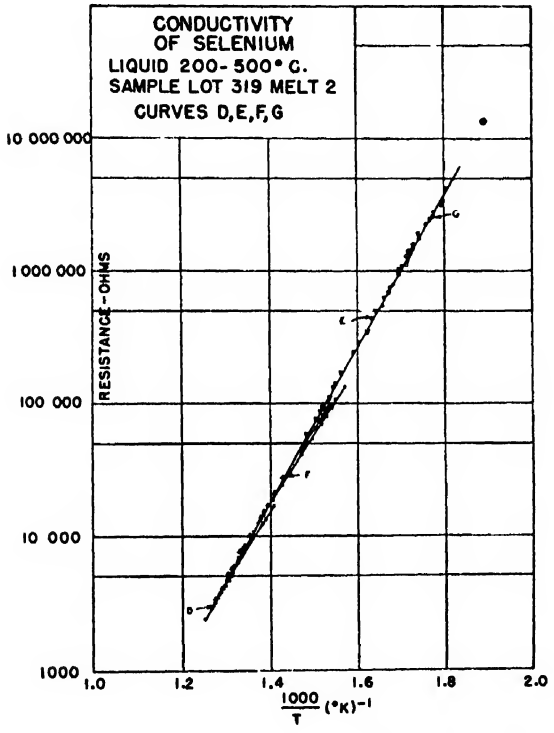


FIG. 3. Resistance of pure selenium, lot 319, melt 2.

contact resistance of metals against selenium and the work function of the metals.[4,5]

Procedure

The cell was carefully washed, rinsed in absolute methyl alcohol and dried in an oven. Selenium shot was introduced and melted. The heating procedure was varied from run to run in order to establish the dependence, if any, of the resistance on previous treatment. The thermocouple e.m.f. was measured simultaneously with the resistance. Care was taken to avoid contaminating the selenium with dust. Any contamination

[4] Schottky, Zeits. f. Physik 118, 539 (1941).
[5] A. V. Ioffe and A. F. Ioffe, "Contact of semiconductors with metals," Bull. Acad. Sci. Ser. Phys. 5, 550 (1941).

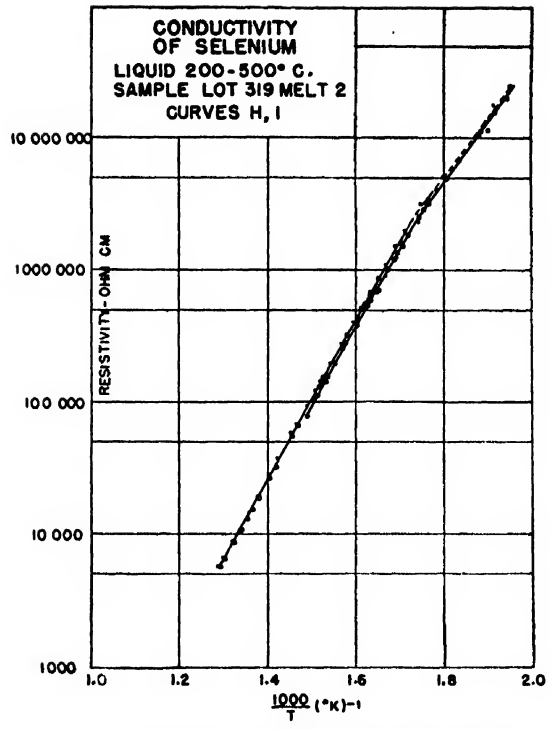


FIG. 4. Resistivity of pure selenium, lot 319, melt 2.

was evidenced by the presence of small crystallized spots in the amorphous selenium surface after cooling slowly to room temperature.

Accuracy of Measurements

The precision of measurement of resistance was limited by the spread of the null indication. At low resistances the error was less than 1 percent, at high values (10 megohms) perhaps as high as 5 percent. The error in reading temperatures was about ±1.00°C arising from difficulty in balancing the bridge and the potentiometers simultaneously. The cell calibration was accurate to ±5 percent.

Data

The data collected is plotted in curves in Figs. 2–12. The results are reviewed in the discussion.

DISCUSSION

Pure Selenium

Care was taken in assembling the data to attain equilibrium conditions. Time rates of change of temperature varied from run to run but averaged about two or three degrees per minute. This rate gave satisfactory curves throughout most of the range. Below 300° a slower rate was desirable. Curves were retraced on reversing the sign of the temperature change. At low temperatures hysteresis loops occurred in a few cases (see curves *N*, *O*, *Q*, *V*, *AE*).

TABLE II.

Sample	Curve	Temp.	A	B
lot 319 melt 2	H, I	215–500	−3.99	5900–6010
lot 319 melt 2 1% I_2	J	215–500	−1.79	3980
I_2 J	K	215–500	−1.75	4020
I_2 K	L	215–500	−1.96	4240
I_2 L	M	215–500	−2.12	4420

TABLE III.

Sample	Curve	Temp.	A	B
lot 319 melt 3	N	250–500	−3.87	5930
lot 319 melt 3 with 0.1% mercury	O	250–500	−3.87	5930
lot 319 melt 3 with 0.1% mercury	Q	250–500	−4.00	6040

Curves $A, B, C, D, E, F, G, I, N, V, W, X, Y, AG, AH$ give data on pure selenium. Temperature dependence of resistance followed the law

$$\log_{10}\rho = A + B/T$$

with values of A and B given in Table I. An average of the values of B gives 5850 with maximum deviation 3.4 percent, of A gives −3.81 with 10 percent deviation.

Curves $H, I, O, Q, V, W, X, Y, AD, AE$, perhaps N, AG and AH show a slight change in temperature coefficient about 300°C. The kink is most easily noticeable in experiments with selenium that has been exposed to air a considerable length of time. This change has been attributed to the presence of SeO_2.

Curve B shows the effect of a short period of heating at about 400°C on the resistance of pure selenium. No change in resistance is apparent. A study of temperature coefficients, B, in Table I, does reveal a slight change in slope with prolonged heating but this is almost within the experimental error of the procedure. The curves AD, AE, AF of Fig. 10 of results on pure melts 5 and 6 give rather low temperature coefficients ($B=5200$ in curve AD). Incidentally, the kink at about 300° is well shown in AE. It was evident from the appearance of crystallized portions in the melt on cooling in the oven that these samples were contaminated. Sample AF completely crystallized on cooling. After the experiment was completed the entire apparatus (oven, hood, etc.) was cleaned thoroughly. Lot 504–2 melt 1 then repeated the values of slope obtained in former samples.

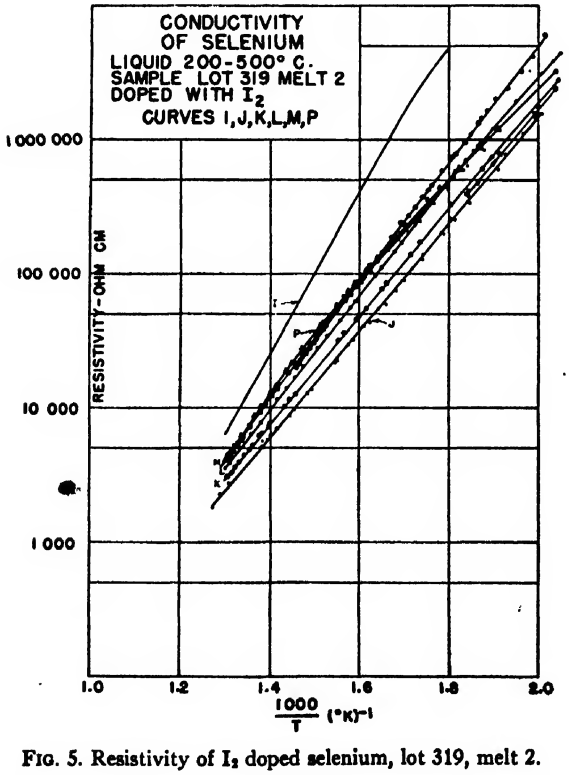


FIG. 5. Resistivity of I_2 doped selenium, lot 319, melt 2.

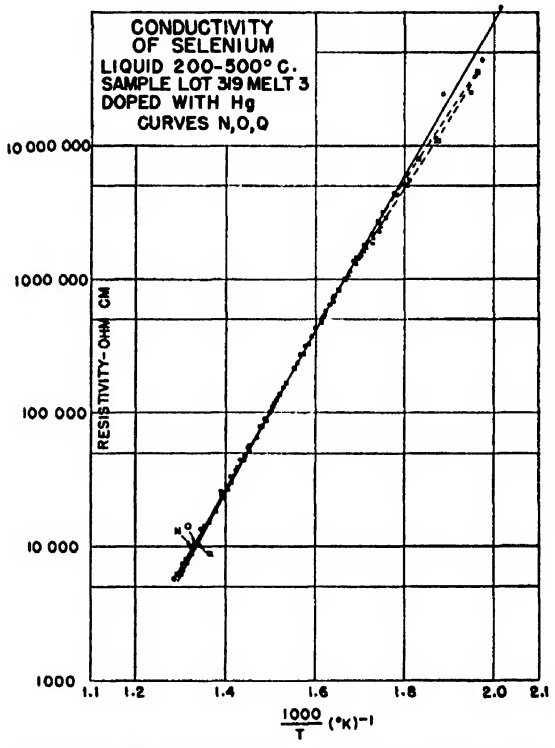


FIG. 6. Resistivity of Hg doped selenium, lot 319, melt 2.

The results of H Pélabon are plotted in Fig. 12. At 300° the resistance decreased suddenly with a velocity such that measurements were not possible. This observation was not checked by our data. At temperatures below 220° all samples crystallized (except those containing phosphorous) with a velocity depending on the amount and type of impurity present.

The data of Borelius is plotted in the same figure. Temperature coefficients varied considerably among samples of different purity (as would be expected) but also changed with different measurements on the same

sample and with depth of the conductivity electrodes in the melt. An equilibrium of two liquid phases of different densities and conductivities was postulated. In our data all pure samples gave the same results. No variation of resistance with depth was noticed. A. N. Campbell and S. Epstein[6] found no indication of shifting internal equilibrium in density measurements over the range 20–277°C. Dobinski and J. Wesolowski[7] found the density exhibited no singularities between 217° and 300°C. The viscosity temperature curve of selenium is given by Dobinski and Wesolowski[8] that of sulfur by

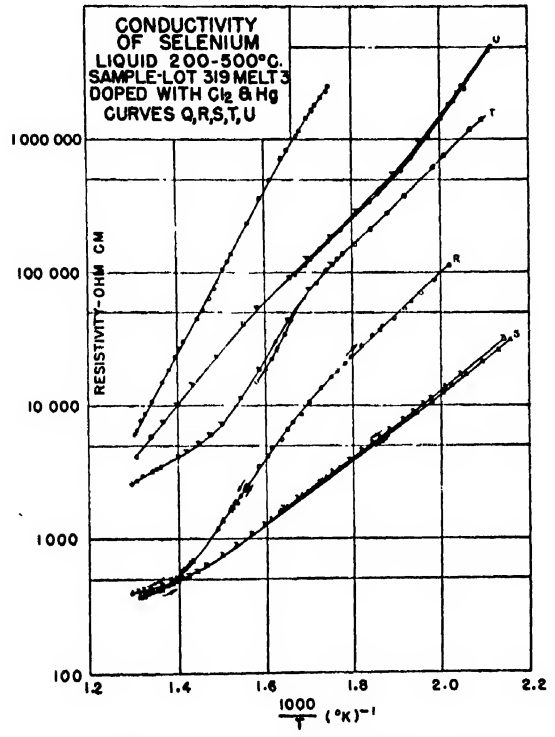


FIG. 7. Resistivity of selenium doped with Hg and Cl_2, lot 319, melt 3.

Bacon and Fanelli.[9] Selenium shows no tendency to form a ring structure at low temperatures as evidenced by the shape of the viscosity curve.

Doped Selenium

The melt of curve *I* was cooled to about 250°C and 1 percent by weight of iodine Se_2I_2(?) was introduced. After a mixing period of one minute, curve *J* was traced. Curve *K* was obtained after a 3½-hour period of heating above 400°C, curve *L* after additional 5 hours, curve *M*

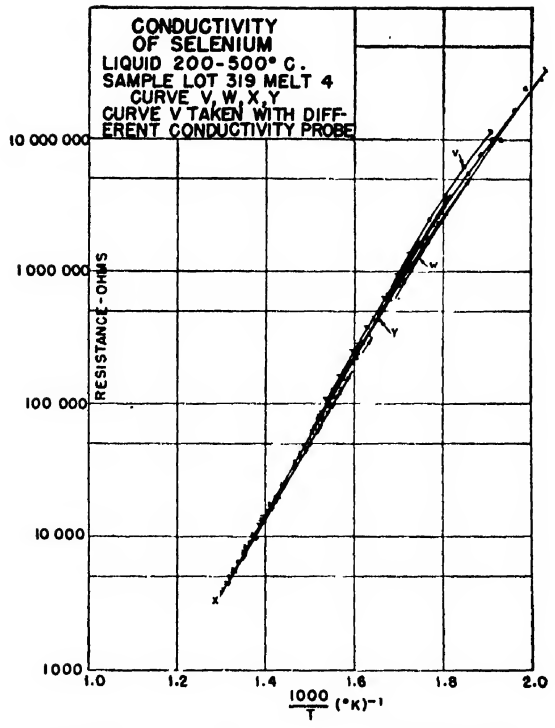


FIG. 8. Resistance of pure selenium, lot 319, melt 4.

after an additional 1¾-hours heating. Table II gives the changes in *A* and *B*. The temperature coefficients are observed to have increased slowly as I_2 was expelled.

0.1 percent mercury as HgSe was added to pure selenium of curve *N* and curve *O* was traced. No change in *A* or *B* was noted. The melt was then heated for 1 hour above 400°C and curve *Q* obtained. Again as indicated in Table III, no definite change in either *A* or *B* was observed.

Melt *M* was then doped with 0.1 percent mercury and immediately retested. Whereas no kinks were evident in curve *M* two appeared in curve *P* (see Table IV). Since conditions were the same in experiments *M* and *P* the changes in temperature coefficients at various temperatures must be attributed to the activity of various compounds of mercury and iodine. Mercury by itself has no effect on temperature coefficients as shown in curves *O* and *Q*. The possible significance of this fact will be discussed later.

The effect on an addition of 1 percent Cl_2 as Se_2Cl_2 to the melt of experiment *Q* is shown in curve *R*. The selenium solidified during addition of Se_2Cl_2 so that run *R* may have been conducted on an inhomogeneous sample. After the melt was heated to 500°C, curve *S* was obtained on cooling. Two temperature coefficients were measured for curve *S*, from 190 to 395°C, 2460, from 395–500°C, 1460. The effect of periods of heating on this melt are shown in Curves *T* and *U*. Different coefficients are noticed in various ranges. Compounds of mercury and chlorine could be active. To check the effect of

[6] A. N. Campbell and S. Epstein, "Density of selenium," J. Am. Chem. Soc. 64, 2679–80 (1942).
[7] S. Dobinski and J. Wesolowski, "Density of selenium," Bull. Acad. Polon. 8–9A, 446–50 (1936).
[8] S. Dobinski and J. Wesolowski, "Viscosity of Liquid Selenium," Bull. Acad. Polon. 49(A), 7–14 (1937).
[9] Bacon and Fanelli, "Viscosity of sulfur," J. Am. Chem. Soc. 65, 639–48 (1943).

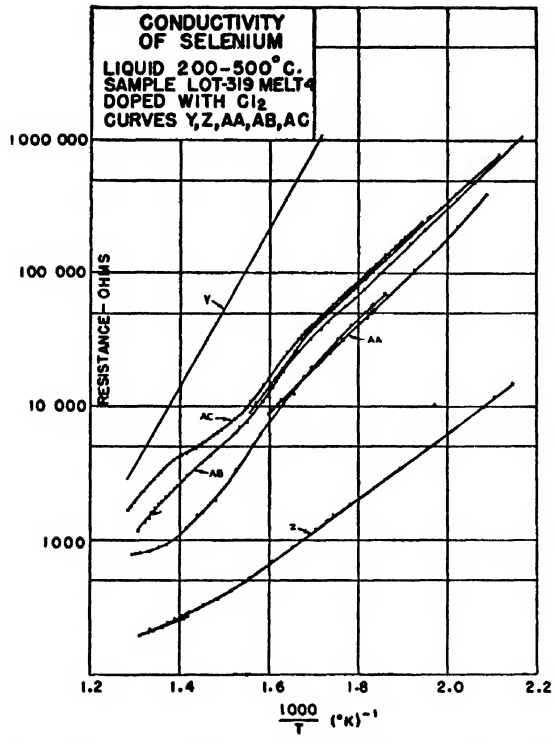


FIG. 9. Resistance of Cl_2 doped selenium, lot 319, melt 4.

chlorine, experiments *Y, Z, AA, AB, AC* were conducted on melts of pure selenium with Cl_2 added as Se_2Cl_2. This time the curve for low slope was obtained immediately. Temperature coefficients were almost identical in curves *S* and *Z*. (See Table V.) (Neither of these melts crystallized on cooling in the oven. This observation is explained in the following manner. Small amounts of impurities act as crystallization centers whereas excessive amounts inhibit the growth of crystals breaking up the continuity of crystallites as in the analogous case of oxygen in molten iron.) Curves *AA, AB, AC* have the same general appearance. In the last two cases a change in temperature coefficient seems to occur about 327°, one at 377° and another at 447°C. Interpretation of the curves is difficult. It is noteworthy that in all cases the resistances of the melts increased steadily with expulsion of chlorine by heating. Various compounds of platinum and chlorine and possibly oxygen influence the temperature coefficient. Experiments were conducted with melts containing 1 percent P. Equilibrium conditions were difficult to attain. A distinct change in temperature coefficient was noted at about 358°C. Below that temperature the value of *B* was 2170, above about 4200. This last figure is not very reliable. It is included to show the direction of the bend and the fact that the original value of *B* was never equaled or exceeded with impurities in the melt. (Melts with phosphorous would not crystallize under any condition. An explanation lies in the netting of molecular

TABLE IV.

Sample	Curve	Temp.	*A*	*B*
lot 319 melt 3 with I_2	*M*	215–500	−2.12	4420
lot 319 melt 3 with I_2 and mercury	*P*	215–275	−0.41	3400
		275–388	−1.05	3760
		388–500	−2.42	4650

TABLE V.

Sample	Curve	Temp.	*A*	*B*
lot 319 melt 3 with 0.1% mercury	*Q*	250–500	−4.00	6040
above with 1% Cl_2	*S*	190–395	−0.84	2460
		395–500	0.54	1460
lot 319 melt 4	*W, X, Y*	250–500	—	5930
above with 1% Cl_2	*Z*	190–395	—	2430
		395–500	—	1460

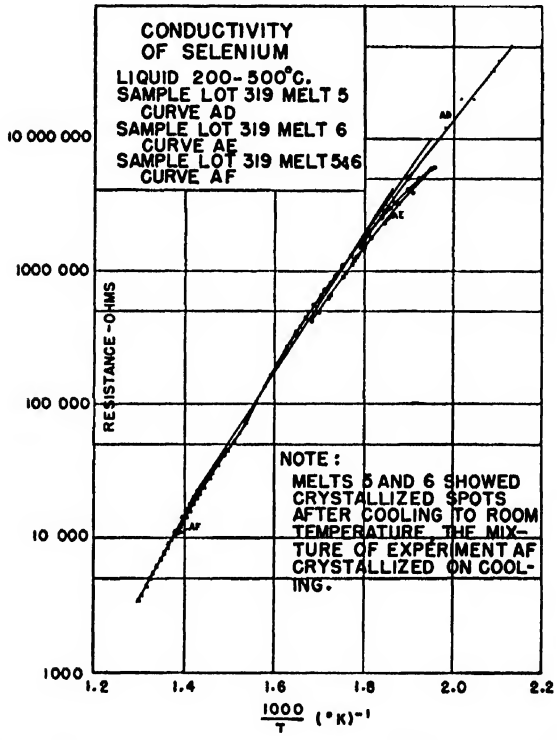


FIG. 10. Resistance of pure selenium, lot 319, melts 5 and 6.

TABLE VI.

Sample	Temperature	*A*	*B*
Pure Selenium	200–500	−3.81	5850
Above with 1% I_2	215–500	−1.79	3980
Above with 1% Cl_2	190–395	−0.84	2430
	395–500	0.54	1460
Above with 1% P	225–358	—	2170
	358–500	—	ave 4240 (?)
Above with 0.1% mercury	250–500	−3.87	5930

threads caused by the trivalence of the element as discussed in Technical Report No. 9.)[1]

Data on different impurities in selenium is collected in Table VI.

The data indicates that selenium when pure, behaves as an ideal semiconductor in the range of temperatures measured. Non-metal impurities decrease the activation energy of the element in an amount depending on the type of impurity. Note the wide difference between the effect of iodine and chlorine, and to a smaller extent on the amount of impurity (experiments H, I, J, K, L, M). The impurity mercury was chosen as one usually considered to have a radical effect on the electrical properties of selenium. Experiments N, O, Q show that mercury was ineffective in changing the activation energies of the pure melt but was active in combining with the non-metals already in the selenium (curve P). Perhaps the discrepancy between the results of Penin and Astakhov[10] and rectifier manufacturers mentioned in Technical Report No. 9 can be resolved by assuming the experiments of the former were carried out on pure selenium whereas selenium doped with halogens was used commercially.

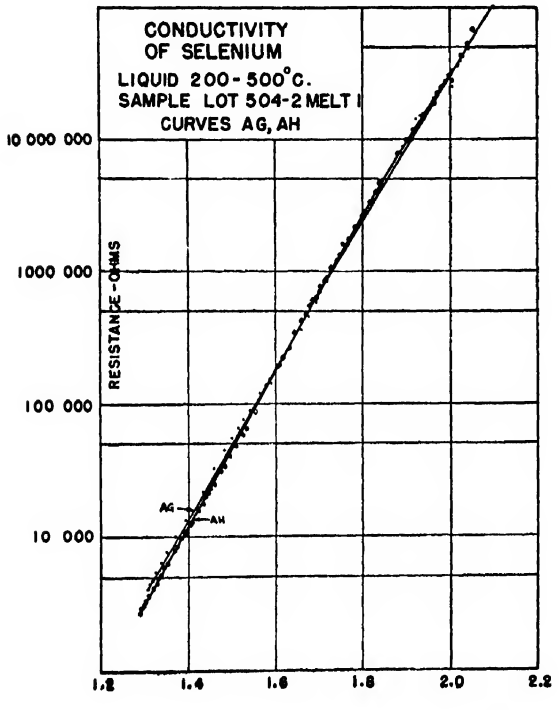


FIG. 11. Resistance of pure selenium, lot 504–2, melt 1.

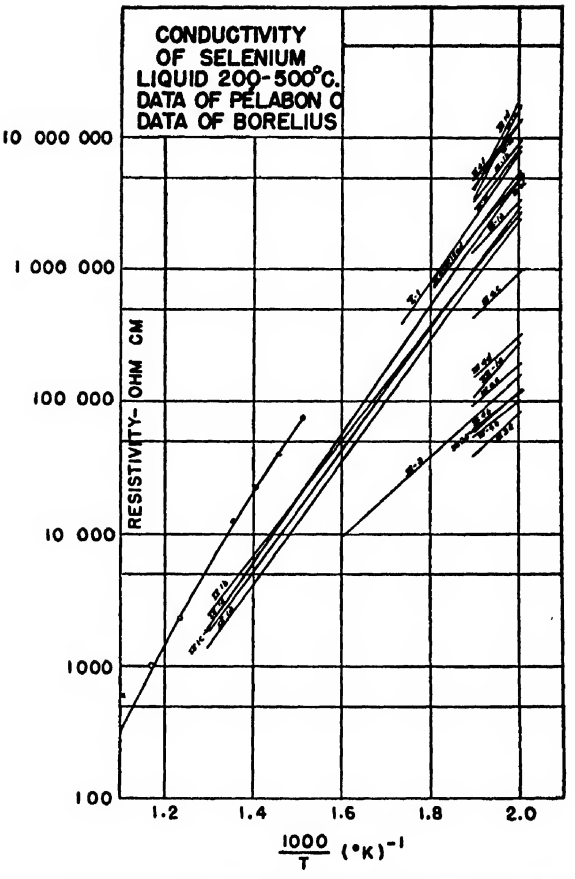


FIG. 12. Resistivity of selenium (data of Pélabon and Borelius).

Melts with metals and non-metals behave in a manner difficult to explain. It is however reasonable to assume that different compounds are stable in different temperature regions and that these compounds should influence the activation energies. There may perhaps be a certain amount of ionic conductivity. However the presence of a large thermoelectric power in liquid selenium and the value of activation energy indicates electronic conduction for the pure semiconductor.

[10] N. A. Penin and K. V. Astakhov, "Effect of admixtures to selenium on the electrical properties of selenium rectifiers," J. Tech. Physics (USSR) 16, 199–206 (1946).

The Specific Heats of MgO, TiO_2, and ZrO_2 at High Temperatures*

JAMES S. ARTHUR
Department of Physics, The University of Pittsburgh, Pittsburgh, Pennsylvania
(Received June 23, 1949)

Mean specific heats were determined from 20°C to 800°C for MgO, TiO_2, and ZrO_2. True specific heats were then calculated assuming that $C_T = A + BT + CT^2$. At 350°C the true specific heats were calculated to be MgO 0.276; TiO_2 0.210; ZrO_2 0.140.

SOME data on the specific heats of oxides in the high temperature range, 20° to 1500°C, are available.[1,2] One of the methods of measuring specific heats employs an adiabatic calorimeter of the type described by Harry Hill and R. M. Bell.[3]

The calorimeter used in obtaining the data listed in this article was patterned after the one used by Hill and Bell with two main changes. An electrolytic, bath type heater was used instead of an electric coil heater in the jacket. The purpose of the heater is to enable the experimenter to match the temperature of the jacket surrounding the calorimeter cup with that of the cup and sample. The electrolyte used was tap water. Advantage is gained in getting more uniform heating and a much shorter reaction time. In addition, the plywood box enclosing the calorimeter cup was lined with aluminum foil to further reduce heat losses.

TABLE I. Mean specific heat of MgO.

Mean specific heat	t_2(°C)	t_1(°C)
Arthur		
0.2425	193.4	23.45
0.2466	197.0	24.25
0.2500	223.8	23.32
0.2519	259.2	25.75
0.2593	389.0	26.35
0.2590	398.7	26.18
0.2590	414.6	30.75
0.2657	483.8	30.08
0.2645	502.1	30.36
0.2733	582.8	31.75
0.2733	594.1	33.15
0.2687	595.2	35.81
0.2775	774.8	35.39
0.2806	800.7	37.00
0.2853	830.4	39.27
Wilkes		
0.2335	100	30.0
0.2470	300	30.0
0.2590	500	30.0
0.2690	700	30.0
0.2765	900	30.0

The calorimeter was calibrated by a small electric heating coil, as described by Hill and Bell.[3]

The heat capacity of the cup was found to increase with temperature in the range 20° to 40°C.

The oxide samples were obtained in powder form and packed into metal foil in the form of cylindrical pellets.[4] The dimensions of the pellets were slightly smaller than the dimensions of the cavity in the calorimeter cup. Aluminum foil was used for temperatures under 600°C and platinum foil when determinations were made at higher temperatures. A mold was used to make the pellets. It consisted of a cylinder composed of two close fitting pieces of hard wood with a brass insert, and a tamping piston made of dowel wood.

A Hoskins electric furnace, type FH204C was used to heat the samples. The temperature was regulated by a Brown electric pyrometer. All high temperature measurements were made with Chromel-alumel or iron-constanton thermocouples.

The methods used to calculate mean specific heat and true specific heat may be found in an article by Wilkes and Wood.[5] Values of mean specific heat are obtained directly from the data recorded and may be written as:

$$C_m = A + B/2(t_2 + t_1) + C/3(t_2^2 + t_2 t_1 + t_1^2), \qquad (1)$$

where C_m = mean specific heat, t_2 = furnace temperature, t_1 = final cup and sample temperature, and A, B, and C

TABLE II. Mean specific heats of TiO_2.

Mean specific heat	t_2(°C)	t_1(°C)
0.1720	123.2	24.10
0.1687	197.1	21.78
0.1748	200.1	24.30
0.1780	228.6	29.31
0.1838	289.2	27.02
0.1899	393.0	28.36
0.1930	400.6	29.56
0.1967	431.2	33.86
0.1959	455.8	31.56
0.1992	487.9	36.05
0.1929	491.8	31.50
0.1976	591.0	33.74
0.1965	593.4	34.12
0.2051	601.4	36.00
0.1967	602.5	32.50
0.2087	799.1	38.40

* From a thesis submitted in partial fulfillment of the requirements for the M.A. degree at Washington and Jefferson College.

[1] G. B. Wilkes, "The specific heats of MgO and Al_2O_3 at high temperatures," J. Am. Ceram. Soc. 15, 72–7 (1932).

[2] Gronow and Schweite, "Die Spezifischen Warmen von CaO, Al_2O_3, $CaO \cdot Al_2O_3$, $3CaO \cdot Al_2O_3$, $2CaO \cdot SiO_2$, und $2CaO \cdot Al_2O_3 \cdot SiO_2$ von 20° bis 1500°C. Zeits. Anorg. allgem. Chem. 216, 185–95 (1933).

[3] Harry Hill and R. M. Bell, "A versatile calorimeter for specific heat determinations," A.S.T.M. Bull. No. 151 March, 1948.

[4] Samples were obtained C. P. from Eimer and Amend.

[5] G. B. Wilkes and O. C. Wood, "the specific heat of thermal insulating materials, heating, piping and air conditioning," A. S. H. V. E. Journal Section, p. 370, June (1942).

TABLE III. Mean specific heats of ZrO_2.

Mean specific heat	t_2(°C)	t_1(°C)
0.1138	195.0	26.82
0.1163	197.4	23.20
0.1188	239.2	25.04
0.1222	278.7	26.62
0.1274	388.1	24.98
0.1275	398.1	29.17
0.1289	430.2	29.38
0.1308	488.9	31.17
0.1289	491.9	28.50
0.1332	579.8	32.56
0.1351	584.7	33.44
0.1332	606.0	32.64
0.1364	633.7	34.58
0.1363	795.7	40.26
0.1351	824.6	37.88
0.1437	943.4	42.87
0.1344	991.8	41.03

are constants. The true specific heat at any temperature is given by:

$$C_T = A + Bt + Ct^2, \qquad (2)$$

where C_T = true specific heat, t = temperature, and A, B, and C are the same constants as used in (1).

DISCUSSION OF RESULTS

The chief function of this article is to list the mean specific heats of MgO, TiO_2 and ZrO_2 in the range 20°–800°C and the true specific heats for the same oxides. The accuracy of the specific heats is judged to be ±2.0 percent.

MgO has a mean specific heat of 0.246 in the range 25°–200°C and 0.282 in the range 37°–800°C (see Table I). TiO_2 has a mean specific heat of 0.175 in the range 24°–200°C and 0.206 in the range 37°-800°C (see Table II).

ZrO_2 has a mean specific heat of 0.115 in the range 24°–200°C and 0.136 in the range 38°–800°C (see Table III).

Table I shows a comparison of the values of the mean specific heat of MgO obtained by the author with those obtained by G. B. Wilkes using an adiabatic bomb calorimeter. The author's values were consistently higher by approximately 0.3 percent for the same temperature range.

TABLE IV. Values of A, B, and C for MgO, TiO_2, and ZrO_2.

	$C_T = A + Bt + Ct^2$		
Material	A	B	C
MgO	0.2342	0.000124	−0.0000000170
TiO_2	0.1512	0.000245	−0.000000222
ZrO_2	0.0988	0.000173	−0.000000156

TABLE V. True specific heats of MgO, TiO_2, and ZrO_2.

Material	Temperature (°C)	True specific heat (Cal/g-deg. C.)
MgO	200	0.258
	350	0.276
	500	0.292
TiO_2	200	0.191
	350	0.210
	500	0.218
ZrO_2	200	0.127
	350	0.140
	500	0.146

Table IV is a list of the values of the constants A, B, and C which determine the equation of true specific heat in terms of temperature:

$$C_T = A + Bt + Ct^2.$$

Some values of C_T are listed in Table V. The values of true specific heat are accurate for temperature ranges approximately half that of the mean specific heats.

Errors resulting from variation in heat capacity of the cup with temperature are estimated at ±2.0 percent.

Radiation errors were minimized in the following ways: (1) A small refractory furnace was used to transfer the hot sample from the large furnace to the cup; (2) a baffle and guide arrangement was used to guide the sample into the cup and prevent direct heat radiation from the small furnace into the cup; (3) rapid transfers of samples were made, and mainly, (4) the use of the adiabatic heating jacket to keep the cup surroundings at the same temperature as the cup.

The Effect of a Video Filter on the Detection of Pulsed Signals in Noise*

DAVID MIDDLETON
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received November 30, 1949)

The effects of inserting a video filter of finite width on the observability of pulsed signals in random noise are examined. It is found that at match (when the pulse and IF filter are each other's conjugate Fourier transforms) no improvement is gained from such a video: the infinitely wide response yields the optimum results. This is strictly true when the second detector is a quadratic rectifier; (a slight improvement on narrowing the video is noted, however, for strong signals when a half-wave linear rectifier is used). Away from match a video filter does give noticeable improvement when the pulses are overlong, i.e., narrower spectrally than the IF ($\lambda>1$); more noise than signal is then removed by the narrower video filter. The greatest gain over the infinite video is observed for final filters slightly wider than the original pulse. On the other hand, for pulses that are too short ($\lambda<1$) the performance is worsened. The mean maximum signal level is decreased more rapidly than the interfering noise background. In all instances the familiar phenomenon of modulation suppression arises. Two different pulse shapes are considered: (1) Gaussian and (2) rectangular, but for wide videos pulse shape is not a critical factor. Curves showing the output signal-to-noise ratios as a function of the input ratio are included for conditions of match ($\lambda=1$) and mismatch ($\lambda\neq1$).

I. INTRODUCTION AND DISCUSSION

IN an earlier paper[1] the factors governing the reception of pulsed (radar) signals in random noise were discussed, and expressions for the output signal-to-noise ratio in terms of the input ratio were derived for three principal modes of perception: (1) observation of an oscilloscope screen, (2) listening to an harmonic of the pulse repetition frequency (PRF) or (3) noting the mean change in a meter's deflection in the presence and absence of a signal. Here, however, we are concerned with visual reception only, and it is the aim of this paper to examine the effects of inserting a finite video filter in the stage before observation (instead of the infinitely wide, uniform response assumed in the earlier analysis),[1] particularly when the pulse and IF filter are not matched, i.e., are not each other's conjugate Fourier transforms. Recent experiments by Fromm[2] indicate good qualitative agreement with the present theory. Some as yet unpublished,** theoretical discussion of the video filter and extensive experimental investigations, supporting our present conclusions were carried out during the war.[3]

The IF and succeeding stages of the receiver have a critical effect on reception. Coming into the IF is a fluctuating noise voltage and also a signal which is a known function of the time. After filtering and rectification the video amplitude cannot be represented as the simple sum of pure signal and "pure" (i.e., normal random) noise, as there are complicated modulation products generated between the two, especially following a linear detector. If the signal is repeated periodically there will be a certain interval during which the amplitude on the oscilloscope screen is on the average a maximum. Our definition, therefore, of what constitutes a signal in the visual case treated here is *the excess of the average of the amplitude, which we designate as* $s_v(0)_G$, *during this interval, over the mean amplitude* ($=n_v$, the r.m.s. video noise voltage) *in the absence of signal.* The concept of the average is needed because sometimes the noise interferes constructively and sometimes destructively with the signal. The above is called the "simple" criterion for visual detection[1] because, in practice, examination merely of the output ratio $s_v(0)_G/n_v$ neglects certain factors. For example, the effects of the (time) widths of the pulse *versus* the widths of the "blades of grass," or noise, are not considered, and in particular our definition does not take into account the action of pulse repetition: the more frequent the pulse, the more easily the eye integrates (in the limited time at its disposal) to detect the mean position of the pulse or pip with respect to the fluctuating background. To test the simple theory, therefore, we need a device, like the peak-reading meter,[2] which does not depend on PRF, but which gives $s_v(0)_G$ directly. The results are applicable to the radar problem within the limitations described above when the eye and brain constitute the final integrating elements. A table listing the various parameters—λ, Ω, Ω', involving the filter widths, pulse widths and durations, etc., that appear in the analysis—is included at the end of this section. Figure 1 illustrates the two types of pulse treated here and their respective spectra.

Figures 2–4 show the output video signal-to-noise ratio as a function of $\sigma_0^2\lambda$, where σ_0 is defined as the ratio of the r.m.s. maximum signal amplitude to the r.m.s. noise amplitude after passage through an IF filter which is matched to the incoming pulse [Eq. (2.13)]; for the Gaussian pulses and filter λ represents the ratio

* The research reported in this document was made possible through support extended Cruft Laboratory, Harvard University, jointly by the Navy Department (ONR) and the Signal Corps, U. S. Army, under Contract N5ori-76, T. O. 1.

[1] J. H. Van Vleck and D. Middleton, "A theoretical comparison of the visual, aural, and meter reception of pulsed signals in the presence of noise," J. App. Phys. 17, 940 (1946).

[2] W. E. Fromm, "An Experimental Investigation of the Reception of Pulsed Signals in the Presence of Noise," Thesis (Polytechnic Institute of Brooklyn, New York, May, 1948).

** *Note added in proof:* A similar analysis of the video filter problem is discussed in section 8.7 of reference 3 above. The results of both researches are in close and satisfactory agreement.

[3] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*, Book No. 24 (Massachusetts Institute of Technology) Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1950.

of IF width to pulse width (frequency), cf. Eq. (2.15) and Table I. The factor λ is included to take into account the increased power necessary to give the same value of s_v/n_v at the output when the pulse and filter are no longer matched ($\lambda \neq 1$). Thus, Figs. 2–4 may be compared directly to show the well-known result[1] that at match a given value of the output ratio s_v/n_v (for the same arbitrary video filter width) is always obtained for the least energy per pulse or, equivalently, for the smallest average pulse power. Figures 2–4 also compare the Gaussian and rectangular pulses of initially the same energy [condition (2.23)]. At match for the Gaussian pulses ($\lambda=1$) the rectangular pulse is slightly off match, since our Gaussian IF is no longer the Fourier transform of the rectangular pulse. For this reason $(s_v/n_v)_{\text{rect.},\lambda=1}$ lies somewhat below the values of $(s_v/n_v)_{\text{Gauss.},\lambda=1}$ for the same video filters (same Ω).

Insertion of a finite video filter *at match* gives no improvement in the Gaussian case, provided the signal is weak compared to the noise, i.e., provided that the linear detector is functioning essentially like a quadratic one.[4] When the signals are strong ($\sigma_0^2 > 10$ db), however, a slight improvement is observed (Fig. 3) for video widths of the order of the original pulse width, i.e., $\Omega \sim 1$, 2. This occurs because a video filter helps to decrease the effects of the low frequency ($n \times n$) noise components arising from the higher-order modulation products generated in the linear detector, products which are absent in the output of the square-law rectifier. Insertion of a video filter away from match, however, will give significant improvement when the pulse is sufficiently long, a fact verified experimentally by Fromm.[2] The gain in performance is particularly noticeable when the IF is much greater than the pulse's spectral spread (see, for example, $\lambda=10$, Fig. 4). Too much noise is then passed by the IF along with the signal, and accordingly narrowing the video helps to eliminate some of the added noise products following rectification. The maximum effect occurs for videos slightly broader than the pulse spectrum. Of course, too pronounced a decrease in video width will eventually cut down the maximum signal level more rapidly than the noise, and the system will lose in effectiveness. Such behavior is observed in Figs. 2–4 as $\Omega \to 0$. On the other hand, for an IF narrower than the pulse (say, $\lambda=0.1$, Fig. 2) no improvement is observed: an infinitely wide video still gives the best results. This is because now the maximum signal strength is decreased more rapidly by the insertion of a video filter than is the r.m.s. noise level.

Figure 5(a) shows how the input power required for a given output ratio ($s_v/n_v=0$ db) varies with the pulse or IF band width, for a variety of video responses. The narrower the video the smaller the value of $\bar{\lambda}(\equiv \omega_v/\omega_b)$ at which the mean input power $\bar{P}$ is a minimum. The explanation lies in the fact that for a fixed IF characteristic the longer pulses have fewer significant components, among which correspondingly more of the pulse's energy is distributed; the narrow videos then remove relatively more noise than signal power in such circumstances, since a larger proportion of pulse energy lies in the low frequency harmonics passed by the video filter. The minimum power required for a given output (here $s_v/n_v=0$ db) occurs at match ($\lambda=1$), as predicted earlier. We note, moreover, that pulse shape is not a very critical factor, except when the video filters are very narrow compared to the pulse. Indeed, according to our "simple" criterion, performance at match (for a fixed PRF) is independent of pulse shape; (actually, there is a small modification because visibility of the pulse is somewhat affected by its widths).[3] Long pulses are no easier to detect than short ones in their respective matched conditions, since we observe merely the net increment in signal, which is quite independent of its duration. This is not true, of course, away from match, as Figs. 2 and 4 show: long pulses are then more easily detected than short ones.

Finally, we call attention to the phenomenon of *modulation suppression*, which is particularly noticeable when the signal is comparable to, and less than, the noise, and is characteristic of all nonlinear devices such as the second detector of our present receiver. For ex-

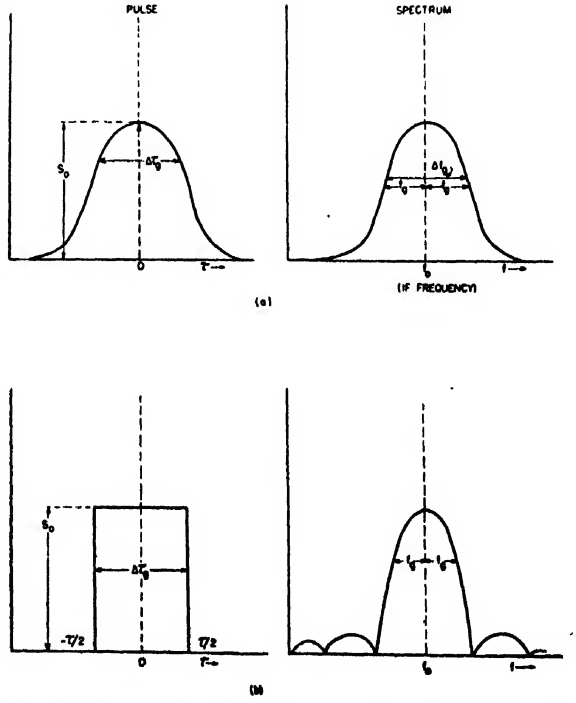


FIG. 1. The wave form and spectrum of (a), a Gaussian pulse, and (b), a rectangular pulse.

[4] This is easily deduced from Eq. (2.16) when $\lambda=1$; only the term $n=0$ in the expression for the noise $\Sigma_n(\Omega')$ (Eq. (2.8)) appears when the detector is a square-law device. We observe that

$$s_v(0)_G/n_v = \sigma_0^2[(\Omega^4+2\Omega^2)/(\Omega^4+2\Omega^2+1)]^{\frac{1}{2}}, \quad (\lambda=1),$$

and clearly the coefficient of σ_0^2 (=a constant) is a maximum when $\Omega \to \infty$, i.e., $\omega_v \to \infty$. This result that the optimum video is one that is indefinitely wide was originally shown by Uhlenbeck (reference 3); see also footnote 19, reference 1, and the footnote following reference 3 above.

ample, when this second detector is the half-wave linear rectifier assumed here, an incoming signal that is weak relative to the noise appears quadratically[5] in the output, instead of linearly; suppression is also noted when the noise is weak, for then the carrier in its turn overwhelms the noise in nonlinear fashion until for sufficiently intense signals the noise becomes relatively ignorable and independent of carrier strength. The envelope, or modulation, is reproduced without distortion. Suppression is evident in Figs. 2–4, where for small signals the slope of the curves is 2, changing to a slope of unity as the input power is increased. Experiment[2] is in good agreement on this point.

II. ANALYSIS

The present theory follows directly from reference 1; we assume explicitly that all filter characteristics are Gaussian. Such responses, while not actually realizeable in practice, are analytically very convenient, and are nevertheless good approximations in most instances,

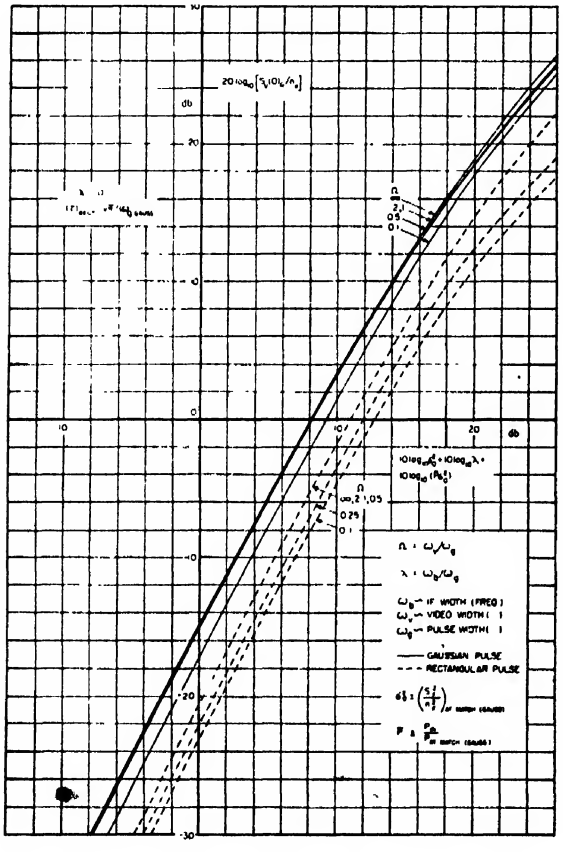


FIG. 2. Output video signal-to-noise ratios for unmatched pulse and IF ($\lambda<1$); same pulse power and peak amplitudes for Gaussian and rectangular pulses entering the IF.

[5] In fact, it has been shown that all half-wave second detectors behave like a quadratic rectifier when the signal is weak compared to the noise. For details see D. Middleton, *Rectification of a Sinusoidally Modulated Carrier in Random Noise*, Proc. I.R.E. 36, 1467 (1948), and Cruft Laboratory Technical Report No. 45, July 1, 1948.

as long as we are not too concerned with frequencies well away from resonance. Two different pulse shapes are considered here: (a) Gaussian and (b) rectangular (see Figs. 1a, 1b), the latter being closer to the particular waveform used in Fromm's work,[2] while the former is a more general shape; pulse shape itself, however, is not a very critical factor in most cases.

Let

$$G_v(\omega)=|G_v(\omega)|e^{-i\omega t_0}, \quad \omega=2\pi f, \quad f>0 \qquad (2.1)$$

be the amplitude response of the inserted video filter; we assume here, and in all our various filters, a phase shift that is linear with frequency—i.e., no phase distortion. The mean-square output noise voltage after rectification is represented by

$$n_v^2=\int_0^\infty W_0(f)df=4\int_0^\infty df\int_0^\infty R_0(t)\cos\omega t dt. \qquad (2.2)$$

Here $W_0(f)$ and $R_0(t)$ are, respectively, the low frequency power spectrum and correlation function after detection, on the assumption of an infinite, uniform video response. If we take account of the finite characteristic of our inserted video filter, (2.2) is modified at once to give

$$n_v^2=\int_0^\infty W_0(f)|G_v(\omega)|^2df=\frac{2}{\pi}\int_0^\infty |G_v(\omega)|^2d\omega \times\int_0^\infty R_0(t)_{(n\times n),\,p=0}\cos\omega t dt. \qquad (2.3)$$

Since the pulsed signal is on only a small fraction of the repetition period, the signal×noise ($s\times n$) products generated in rectification are negligible, and the noise ×noise ($n\times n$) contributions alone are significant; furthermore, the ($n\times n$) noise is essentially unaffected by the presence of such a signal, so that we can safely replace $R_0(t)$ in (2.2) above by $R_0(t)_{(n\times n),\,p=0}$, which is the output correlation function for rectified noise without a signal ($p=0$). The $d-c$ term, corresponding to the constant part of $R_0(t)$, is excluded, since our video filters do not pass the steady component of the detected wave.

For the effect of the video filter on the signal we may write

$$s_v(t)_G=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{i\omega t}S_v(\omega)G_v(\omega)d\omega, \qquad (2.4)$$

where $S_v(\omega)$ is the Fourier transform of the unfiltered video output, given by $[s_v(t)_{p>0}-s_v(t)_{p=0}]$. For real signals $G_v(\omega)$ is required to be equal to $G_v(-\omega)^*$; $s_v(t)_{p>0}$ and $s_v(t)_{p=0}$ correspond respectively to the r.m.s. increments in video signal (infinite, uniform video filter) when there is a pulse and noise and when there is noise alone. The difference of these two quantities is consequently the mean net increment in the video level due

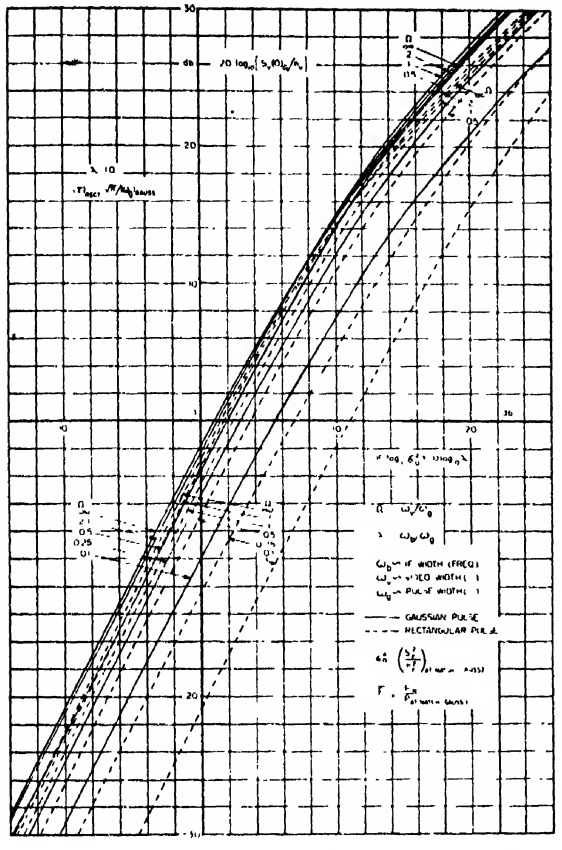


FIG. 3. Output video signal-to-noise ratios for matched pulse and IF; equal pulse power and peak amplitudes for Gaussian and rectangular pulses entering the IF.

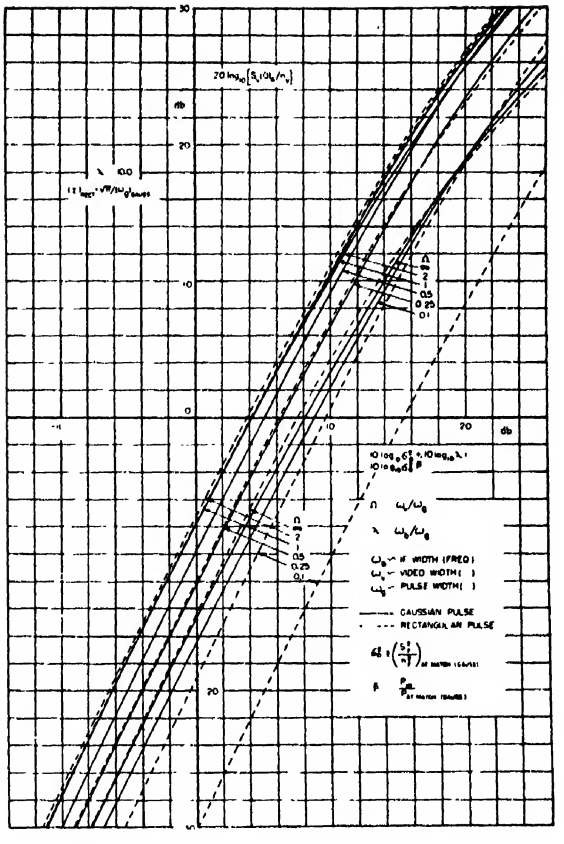


FIG. 4. Output video signal-to-noise ratios for unmatched pulse and IF ($\lambda>1$); same pulse power and peak amplitudes for Gaussian and rectangular pulses entering IF.

to the presence of a signal, corresponding to our definition in Section 1. (The notation here and in what follows is that of reference 1.) With the aid of (2.1) in (2.4) we obtain $s_v(t)_G$, observing that according to our "simple" criterion of visibility we want now the maximum value of $s_v(t)_G$; this corresponds to the maximum of the r.m.s. increment due to signal and noise over the surrounding noise background and is found by setting $t=t_0$, i.e., by selecting our origin on the time-scale so that $t=t_0$ gives a maximum, a gating operation performed automatically by the eye. We have, finally,

$$s_v(0)_G=[s_v(t)_G]_{\max}=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega|G_v(\omega)|$$

$$\times\int_{-\infty}^{\infty}[s_v(t')_{p>0}-s_v(t')_{p=0}]e^{-i\omega t'}dt'. \quad (2.5)$$

Let us now calculate some specific examples. We choose $G_F(\omega)$ and $G_v(\omega)$ to be respectively the amplitude filter responses of the IF and video stages of our receiver and represent their moduli by

$$\begin{aligned}|G_F(\omega)|&=e^{-\omega^2/2\omega_b^2}, \quad \omega_b=2\pi f_b,\\ |G_v(\omega)|&=e^{-\omega^2/2\omega_v^2}, \quad \omega_v=2\pi f_v.\end{aligned} \quad (2.6)$$

In both the above instances we have assumed a linear phase dependence with frequency. The various quantities f_b, f_v, etc., are described in Table I.

The low frequency correlation function for the noise following a half-wave linear detector is found to be [cf. Eq. (24), reference 1] exclusive of the steady component,

$$R_0(t)_{(n\times n),\,p=0}=\frac{\beta^2\psi}{8\pi}\sum_{n=0}^{\infty}\mu_n{}^2r_0(t)^{2n+2} \quad (2.7)$$

where

$$\mu_n=\frac{(2n)!}{2^{2n}n!(n+1)!}=\frac{(1/2)_n}{(2)_n},$$

$$r_0(t)=\exp(-\omega_b{}^2t^2/4); \quad \psi=n_F{}^2=\frac{W_0\omega_b}{2\pi^{\frac{1}{2}}}.$$

The quantities W_0 and ψ are the maximum spectral intensity of the noise entering the IF and the mean noise power output of the IF, respectively; $r_0(t)$ is therefore the (normalized) correlation function of the random noise leaving the IF stage; β is a constant (e.g., the dynamic transconductance) of the detector.

Substituting (2.1) and (2.7) into (2.3) and performing the integration yield the mean-square video noise

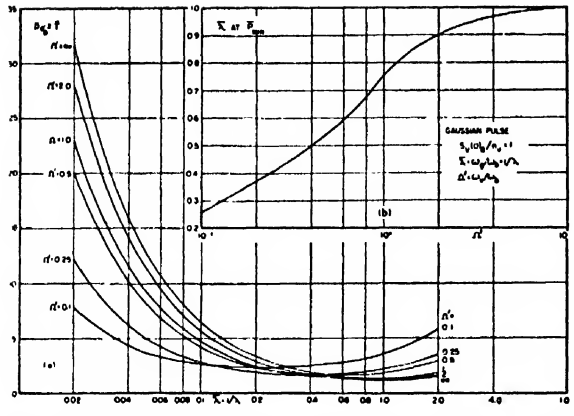


FIG. 5. (a) Average pulse power as a function of pulse length or IF width; (b) minimum average pulse power as a function of the video filter width.

voltage

$$n_v^2=\frac{\beta^2\psi}{8\pi}\sum_n(\Omega')$$

$$=\frac{\beta^2\psi}{8\pi}\sum_{n=0}^{\infty}\mu_n^2\frac{1}{(1+2(n+1)\omega_b^2/\omega_v^2)^{\frac{1}{2}}};$$

$$\Omega'\equiv\omega_v/\omega_b. \tag{2.8}$$

Observe that when the video is infinitely wide, i.e., $\omega_v\to\infty$, the series can be summed to give the familiar value $n_v^2=0.0435\beta^2\psi$, which is our earlier result, Eq. (25), reference 1. Note also that for vanishingly narrow videos ($\omega_v\to 0$) the output vanishes, as one would expect.

For the half-wave linear rectifier considered here, the increment in video signal output is[1,6]

$$s_v(t')_{p>0}-s_v(t')_{p=0}$$

$$=\frac{\beta\psi^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}[{}_1F_1(-\tfrac{1}{2};1;-s_F(t')^2/\psi)-1], \tag{2.9}$$

where ${}_1F_1$ is a confluent hypergeometric function and $s_F(t')$ is the r.m.s. envelope of the carrier after passage through the IF filter. In terms of the original r.m.s. modulation $s_I(t)$ entering the receiver we obtain for $s_F(t')$

$$s_F(t')=\frac{1}{2\pi}\int_{-\infty}^{\infty}G_F(\omega)e^{i\omega t'}d\omega\int_{-\infty}^{\infty}s_I(t)e^{-i\omega t}dt. \tag{2.10}$$

Before considering particular examples we remark that (2.9) is more readily handled if we approximate the hypergeometric function by $x/2-\gamma x^{\frac{1}{2}}$; ($\gamma=0.0763$, $0\leq x\leq 10$) in the manner of reference 1 (p. 962).

Case I. Gaussian Pulse

Here we choose

$$\sqrt{2}s_I(t)=S_0\exp(-\omega_g^2t^2/2), \tag{2.11}$$

where $S_0\equiv$peak value of the envelope and $\omega_g=2\pi f_g$. Then from (2.10) and (2.6) we obtain[7]

$$s_F(t')=s_F\exp(-t'^2b^2/4);\quad b\equiv\sqrt{2}\omega_b\omega_g/(\omega_b^2+\omega_g^2)^{\frac{1}{2}};$$
$$s_F=S_0b/2\omega_g. \tag{2.12}$$

Substituting our results (2.12) into (2.9), with the aid of the approximation for ${}_1F_1$, and applying this to (2.5) enables us to write finally for the maximum average signal

$$s_v(0)_G=\frac{\beta\psi^{\frac{1}{2}}p}{2(2\pi)^{\frac{1}{2}}(1+b^2/\omega_v^2)^{\frac{1}{2}}}$$
$$\times\{1-\gamma'p^{\frac{1}{2}}(1+b^2/\omega_v^2)^{\frac{1}{2}}/(\tfrac{2}{3}+b^2/\omega_v^2)^{\frac{1}{2}}\};$$
$$\gamma'=\gamma(8/3)^{\frac{1}{2}}=0.1246,\quad p\leq 10. \tag{2.13}$$

The quantity p is defined as the mean-square maximum signal output of the IF divided by the mean-square noise amplitude, *viz.*: $p\equiv s_F(t)_{\max.}^2/n_F^2=s_F(0)^2/n_F^2=s_F^2/\psi$. Again we may check (2.13) by noting that as the video is made very wide our signal $s_v(0)_G$ approaches the earlier form, Eq. (25) of reference 1. The desired signal-to-noise ratio $s_v(0)_G/n_v$ follows at once from (2.13) and (2.8). We see from (2.13) that for weak signals the output signal power is proportional to the *square* of the input carrier-to-noise ratio p, an example of modulation suppression discussed in the preceding section. On the other hand, strong signals suppress the noise, so that when $p^2\gg 1$ we obtain from the asymptotic development of ${}_1F_1$ (see Eq. (A3.3) of reference 6) the average video increment

$$s_v(0)_G\backsimeq\frac{\sqrt{2}}{\pi}\beta(\psi p)^{\frac{1}{2}}(1+b^2/2\omega_v^2)^{-\frac{1}{2}},\quad p>20, \tag{2.14}$$

which is now directly proportional to $(p)^{\frac{1}{2}}$. The two expressions (2.13) and (2.14) for the video signal enable us to cover the entire range of values of $s_v(0)_G/n_v$ *vs.* p with a small region of interpolation for $10<p<20$.

At this point it is convenient to introduce a number of parameters

$$\lambda=\omega_b/\omega_g;\quad\Omega=\omega_v/\omega_g;\quad\Omega'=\omega_v/\omega_b=\Omega/\lambda=\Omega\bar{\lambda}. \tag{2.15}$$

The signal-to-noise ratio becomes finally [see (2.8)]

$$\left[\frac{s_v(0)_G}{n_v}\right]_{\text{Gauss}}=\frac{p(1-0.1246(p)^{\frac{1}{2}}[1+2\lambda^2/\Omega^2(1+\lambda^2)]^{\frac{1}{2}}/[\frac{2}{3}+2\lambda^2/\Omega^2(1+\lambda^2)]^{\frac{1}{2}})}{[1+2\lambda^2/\Omega^2(1+\lambda^2)]^{\frac{1}{2}}\{\sum_n(\Omega')\}^{\frac{1}{2}}},\quad p\leq 10, \tag{2.16}$$

[6] D. Middleton, "Some general results in the theory of noise through nonlinear devices," Quart. App. Math. 5, 445 (1948), Section 4 and Appendix A3.

[7] There is a misprint in the expressions for s_F in Eq. (61b) of reference 1; $\sqrt{2}$ should be replaced by 2.

TABLE I.

Pulse	Amplitude of wave form	Power spectrum	Time width* $(\Delta t)_g$	Spectral width* $(\Delta f)_g$	s_F^2/n_F^2	P_{in}
1. Gauss	$S_0 \exp(-\omega_g^2 t^2/2)$	$(S_0^2\pi/2\omega_g)\exp(-\omega^2/\omega_g^2)$	$\frac{(\log_e 2)^{\frac{1}{2}}}{\pi f_g} = 0.265/f_g$	$2(\log_e 2)^{\frac{1}{2}} f_g = 1.666 \cdot f_g$	$\frac{(\pi)^{\frac{1}{2}} S_0^2 \omega_b}{W_0(\omega_g^2+\omega_b^2)}$	$\frac{S_0^2(\pi)^{\frac{1}{2}}}{2\omega_g T}$
2. Rectangular	S_0, $-\tau/2<t<\tau/2$; 0, $\lvert t\rvert > \lvert\tau/2\rvert$	$4S_0^2 \frac{\sin^2(\omega\tau/2)}{\omega^2}$	τ	$2.782/\pi\tau = 0.866/\tau$	$\frac{S_0^2\tau}{2T}$	$(\pi)^{\frac{1}{2}} S_0^2 \frac{\Theta^2\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)}{W_0\omega_b}$
Filter						
3. IF (centered about IF freq. f_0)		$\exp(-\omega^2/\omega_b^2)$		$(\Delta f)_F = 2(\log_e 2)^{\frac{1}{2}} f_b$		
4. Video		$\exp(-\omega^2/\omega_v^2)$		$(\Delta f)_v = (\log_e 2)^{\frac{1}{2}} f_v$		

* The widths of the filters and pulses are measured between half-power (or intensity) points—i.e., points at which the intensity is one-half its maximum value.

Ω = [video filter width (freq.)]/[½ pulse width (freq.)]; $\Omega' = \Omega/\lambda = \Omega\bar{\lambda}$.
λ = [IF filter width (freq.)]/[pulse width (freq.)] $= \bar{\lambda}^{-1}$; T = PRF; $\tau \ll T$.

and for strong signals

$$\left[\frac{s_v(0)_G}{n_v}\right]_{\text{Gauss}} \simeq 4(p/\pi)^{\frac{1}{2}}\left\{\left[1+\frac{\lambda^2}{\Omega^2(1+\lambda^2)}\right]\left[\sum_n (\Omega')\right]\right\}^{-\frac{1}{2}}, \quad p>20. \quad (2.17)$$

Figure 3 illustrates $s_v(0)_G/n_v$ when $s_F/n_F = (p)^{\frac{1}{2}}$ is chosen as independent variable in the matched condition $\lambda = 1(\omega_b = \omega_g)$. The effect of the different videos is observed through the quantity Ω. When the pulse and IF filter are not matched ($\lambda \neq 1$) we may represent the ratio p by

$$p = [2\lambda^2/(1+\lambda^2)]\sigma_0^2, \quad (2.18)$$

where σ_0^2 is taken to be $(s_F/n_F)^2$ at match. In terms of the input power to the IF, which is simply the (energy/pulse)$\times$(number of pulses/sec.), we may also write [cf., Eq. (68), reference 1),

$$p = [2\lambda/(1+\lambda^2)]\sigma_0^2 \bar{P}\bar{T}, \quad (2.19)$$

in which $\bar{P} = P_{in}/P_0$ and $\bar{T} = T_{in}/T_0$, the subscript (0) referring to the input power to the IF (P_0) and pulse repetition period (T_0) at match, while P_{in} and T_{in} are these quantities for arbitrary conditions of operation, not necessarily at match. Figures 2 and 4 show $s_v(0)/n_v$ at mismatch, and here we have chosen as our independent variable $\sigma_0^2\lambda$, or $\sigma_0^2\bar{P}\bar{T}$, in order to take into account the added power input necessary to achieve a given output when away from match.

We mention briefly two additional limiting cases: for a fixed pulse width and very wide IF ($\omega_b \to \infty$, $\lambda \to \infty$; $\Omega' \to 0$, $\Omega = \Omega$) s_v/n_v vanishes as λ^{-1}, since $\sigma_0^2 \sim \omega_b^{-1}$, whereas for a fixed IF and a very narrow pulse (in time) ($\omega_g \to \infty$, $\lambda \to 0$, $\Omega' = \Omega'$; $\Omega \to 0$) we find at once that $s_v/n_v \sim \lambda^2$. Thus in both instances the output vanishes; in the former because of the added noise, in the latter because of vanishing signal strength. Similar remarks apply for the rectangular pulse treated below.

Case II. Rectangular Pulse

Here the shape of the signal wave entering the IF is

$$\sqrt{2}s_I(t) = S_0, \quad -\tau/2<t<\tau/2; \quad =0, \quad |t|>|\tau/2|. \quad (2.20)$$

S_0 is again the peak amplitude, and τ measures the duration of the pulse between half-power points. The various IF and video filters are Gaussian as before (cf. Eq. (2.6)]. The envelope of the pulse leaving the IF follows immediately from (2.10) and the result

$$\int_0^\infty \frac{\sin ax}{x}\exp(-bx^2/2)dx = \pi\phi^{(-1)}(a/b^{\frac{1}{2}}) = \frac{\pi}{2}\Theta[a/(2b^{\frac{1}{2}})], \quad (2.21a)$$

where

$$\Theta(x) = \frac{2}{\pi^{\frac{1}{2}}}\int_0^x e^{-t^2}dt \quad (2.21b)$$

is the well-known error function. We have finally

$$s_F(t') = s_F\{\Theta([t'+\tau/2]\omega_b/\sqrt{2}) - \Theta([t'-\tau/2]\omega_b/\sqrt{2})\}/2\Theta(\tau\omega_b/2\sqrt{2}), \quad (2.22)$$

in which

$$s_F = S_0\Theta\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)\Big/\sqrt{2}. \quad (2.22a)$$

In terms of our definition for p we see that

$$p_{\text{rect.}} = (\pi)^{\frac{1}{2}}S_0^2\Theta(\tau\omega_b/2\sqrt{2})^2/W_0\omega_b = 2\Theta\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)^2\sigma_0^2. \quad (2.22b)$$

Compare this with Eq. (2.18) which gives p_{Gauss}; if we make such a comparison on the basis of equal energy/pulse into the IF, i.e.,

$$E_{\text{Gauss}} = E_{\text{rect.}} = \int_{-\infty}^{\infty} s_I(t)^2 dt,$$

the relation between τ and ω_g is found from (2.11) and (2.20) to be

$$\tau=(\pi)^{\frac{1}{2}}/\omega_g. \tag{2.23}$$

Then we have

$$p_{\text{rect.}}=\left(\frac{1+\lambda^2}{\lambda^2}\right)\Theta\left(\frac{\pi^{\frac{1}{2}}\lambda}{2\sqrt{2}}\right)^2 p_{\text{Gauss}}, \quad (\tau=\pi^{\frac{1}{2}}/\omega_g). \tag{2.24}$$

Thus we are able to discuss the case of rectangular pulses in terms of the parameters (2.15) of the wholly Gaussian system.

With (2.22) for $s_F(t')$ we now use our approximation for ${}_1F_1$ but omit the term containing $s_F(t')^3$, as it leads to analytical difficulties out of proportion to the results achieved; our final expression is not seriously affected, since with the aid of the asymptotic development for ${}_1F_1$ we can interpolate with little error in the region not covered by the small-signal result involving $s_F(t')^2$ alone ($p<1$) and so obtain s_v/n_v as a function of $p^{\frac{1}{2}}$ (or $\sigma_0(\lambda)^{\frac{1}{2}}$) over the entire range of these variables; the parallel behavior with the Gaussian pulse aids the interpolation.

The increment in video signal (2.9) is accordingly[8]

$$s_v(0)_G\doteq p\frac{\beta\psi^{\frac{1}{2}}\omega_v}{16\pi}\int_{-\infty}^{\infty}dt'e^{-\omega_v^2t'^2/2}\{\Theta[(t'+\tau/2)\omega_b/\sqrt{2}]$$

$$-\Theta[(t'-\tau/2)\omega_b/\sqrt{2}]\}^2/\Theta^2\left(\frac{\tau\omega_b}{2\sqrt{2}}\right), \quad p<\tfrac{1}{2}. \tag{2.25}$$

The integrand in (2.25) may be evaluated in straightforward manner if we use a contour integral representation of the error function and the related $\phi^{(j)}(x)$, which are specifically

$$\phi^{(j)}(x)=\frac{d^{(j)}}{dx^{(j)}}\frac{e^{-x^2/2}}{(2\pi)^{\frac{1}{2}}}, \quad j=0, 1, \cdots,$$

and are related to the Hermitian polynomials $H_j(x)$ by $(-1)^jH_j(x)e^{-x^2/2}/(2\pi)^{\frac{1}{2}}=\phi^{(j)}(x)$; the $\phi^{(j)}$ are tabulated up to $j=6$ in T. C. Fry's *Probability and Its Engineering Uses* (D. Van Nostrand Company, Inc., New York, 1928), p. 456.

We obtain finally for the video increment[9]

$$s_v(0)_G\doteq\frac{p\beta\psi^{\frac{1}{2}}}{2(2\pi)^{\frac{1}{2}}}\sum_{n=0}^{\infty}\frac{4}{(2n)!(\Omega'^2+1)^{2n}}$$

$$\times\phi^{(2n-1)}\left(\frac{\tau\omega_v}{2(\Omega'^2+1)^{\frac{1}{2}}}\right)^2\Big/\Theta\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)^2, \quad p\leq\tfrac{1}{2}. \tag{2.26}$$

Compare this with the leading term of (2.13) for the Gaussian pulse. Notice that for an infinitely wide video response ($\omega_v\to\infty$), $s_v(0)_G$ reduces to $s_F^2\beta/2(2\pi\psi)^{\frac{1}{2}}$, as we would expect. Pulse shape is no longer significant relative to the video; all harmonics of the envelope are passed when $\omega_v\to\infty$ and only the energy/pulse is important.

The output signal-to-noise ratio for the rectangular pulse is obtained from (2.8) and (2.26) and is

$$\left[\frac{s_v(0)_G}{n_v}\right]_{\text{rect.}}\doteq p\left\{\sum_{n=0}^{\infty}4\phi^{(2n-1)}\left(\frac{\tau\omega_v}{2(\Omega'^2+1)^{\frac{1}{2}}}\right)^2\right.$$

$$\left.\div\Theta^2\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)(2n)!(\Omega'^2+1)^{2n}\right\}\Big/(\textstyle\sum_n(\Omega'))^{\frac{1}{2}},$$

$$p<\tfrac{1}{2}. \tag{2.27}$$

The noise term is unchanged, since only ($n\times n$) products ($p=0$) enter effectively, and the amount of noise power available in the low frequency spectra region after our linear rectifier then depends on IF filter shape alone, cf. (2.7) and (2.8). Again with the help of the contour integral representation of the error function and $\phi^{(j)}(x)$, we find that for large carrier strengths the video signal is

$$[s_v(0)_G]_{\text{rect.}}\backsimeq\frac{\sqrt{2}\beta s_F}{\pi}\Theta\left(\frac{\tau\omega_v}{2(2\Omega'^2+2)^{\frac{1}{2}}}\right)\Big/\Theta\left(\frac{\tau\omega_b}{2\sqrt{2}}\right);$$

$$s_F^2/\psi>20, \tag{2.28}$$

analogous to (2.14). The signal-to-noise ratio is therefore

$$[s_v(0)_G/n_v]_{\text{rect.}}$$

$$\backsimeq 4(p/\pi)^{\frac{1}{2}}\Theta\left(\frac{\tau\omega_v}{2(2\Omega'^2+2)^{\frac{1}{2}}}\right)\Big/\Theta\left(\frac{\tau\omega_b}{2\sqrt{2}}\right)(\textstyle\sum_n(\Omega'))^{\frac{1}{2}},$$

$$p>20, \tag{2.29}$$

a result to be contrasted with (2.17). Figures 2–4 also give (2.27) and (2.29).

ACKNOWLEDGMENT

The author wishes to express his appreciation to Miss Marilyn Lang and Mrs. Roger Stokey, who performed most of the computations in this paper.

[8] p is written p_{Gauss} or $p_{\text{rect.}}$ only when it is not clear from the context which is meant.

[9] The details of the evaluation are given in Technical Report No. 57, Cruft Laboratory, Harvard University (August 10, 1948).

Negative Wire Corona at High Temperature and Pressure

L. R. Koller and H. A. Fremont
Research Laboratory, General Electric Company, Schenectady, New York
(Received December 27, 1949)

The corona characteristics of air and methyl chloride were studied in a modified type of Cottrell precipitator at temperatures from room temperature to 500°C and pressures from 1 to 5 atmospheres. Within this range the corona current is a power function of voltage with gas density as a parameter. This power is independent of density. Values are given for air and methyl chloride.

INTRODUCTION

In the course of the development of a high temperature, high pressure electrostatic precipitator, it was necessary to determine the corona characteristics of air and methyl chloride. These studies were made in a modified type of Cottrell precipitator at temperatures from room temperature to 500°C and pressures from 1 to 5 atmospheres. The wire sizes were varied from 0.005″ to 0.02″ in diameter. Negative wire corona only was investigated.

APPARATUS

A drawing of the precipitator used is shown in Fig. 1. It was originally constructed for the removal of finely divided powder from a process gas stream. It consisted essentially of a wire mounted along the axis of a 2″ diameter tube. The latter was of seamless steel 36″ long and 2″ in diameter. Great care was used in the construction of the unit to remove sharp edges and roughened surfaces.

A drawn tungsten wire was used for the central electrode. It was not given any cleaning or polishing treatment. No significant pitting or corrosion of the wire was observed during the tests.

The wire was mounted in a bayonet-like fixture at the bottom and spring loaded by means of a tungsten spring at the top. The ends were fastened by ball and socket joints formed by fusing the ends of the wire to small balls about 50 mils in diameter. This eliminated any sharp projecting ends of wire. The purpose of the bayonet fixture was to leave the bottom part of the equipment free for both visual observation of the wire during operation, and also to simplify removal of any powder precipitated when the unit was used as a precipitator.

Zirconium silicate insulators similar to submarine insulators were supplied by the Locke Insulator Company. The leakage over these insulators was negligible in the range of currents investigated.

The unit was wrapped with nichrome wire and insulated with pipe lagging so that it could be heated electrically. Temperatures were determined by three chromel alumel thermocouples peened into the metal and equally spaced along the 2″ pipe.

The electrical circuit is shown in Fig. 2. It consisted of a voltage doubler capable of supplying 25×10^{-3} amps at 40 kv. The positive terminal of the power supply was connected to the precipitator shell which was grounded, and the negative side was connected to the wire. The voltage was controlled by means of a variac in the primary of the high voltage transformer. Voltage (average) was measured on a voltmeter consisting of a microammeter with a 30 megohm series resistance. In order to limit the current in case of arc over an 8013A kenetron, with its emission limited by operating the filament at reduced temperature was included in the high potential side of the line.

The precipitator was assembled with piping containing a flow indicator, pressure gauge and throttle valve. With this arrangement it was possible to introduce various gases at pressures up to 5 atmospheres and make measurements with a small steady flow of gas through the tube. The gas flow was introduced at the upper end of the precipitator as shown in Fig. 1.

PROCEDURE AND PRELIMINARY TESTS

Preliminary tests were made in air at 1 atmosphere. For these tests the end of the unit was sealed with a

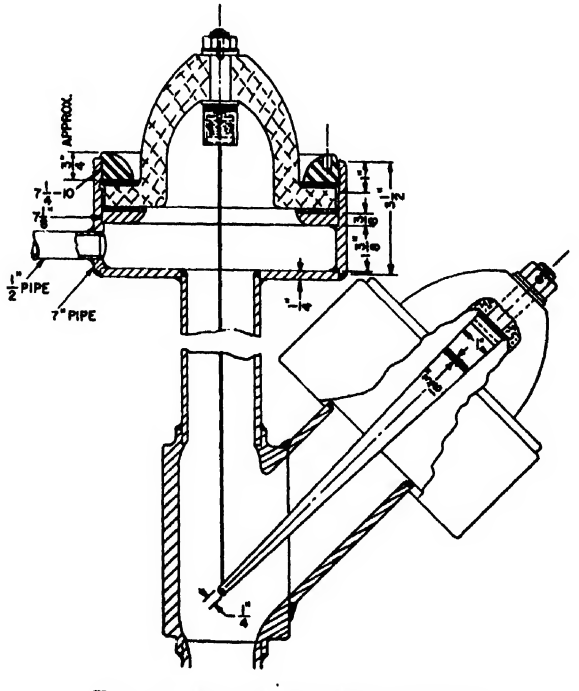


Fig. 1. Design of electrostatic precipitator.

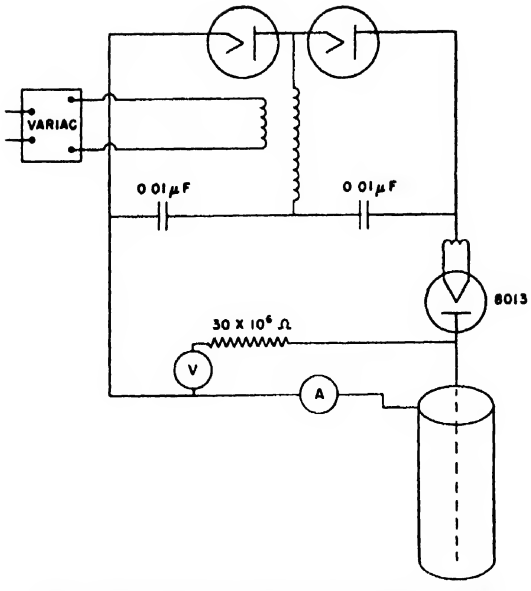


FIG. 2. Electrical circuit for corona measurements.

sheet of $\frac{1}{8}$" transparent cellulose acetate. This made it possible to observe the corona visually by looking in through the end of the tube along the wire.

Voltage was applied to the precipitator and gradually increased while the wire was observed through the open end of the tube. Corona was first noticed around 6000 volts and increased in intensity with increasing voltage. It appeared to be quite uniform along the wire, and there was no tendency to spark or arc over at the ends of the wire or over the insulators. As the voltage was increased still further, the wire was seen to vibrate. At 30 kv, corresponding to a current of about 10 ma spark-over generally took place near the middle of the wire. This may have been due to the close approach to the wall as a result of the vibration.

RESULTS

Typical volt ampere characteristics at 1 atmosphere and room temperature are shown in Fig. 3 for air, for a mixture of air and approximately 19 percent of CCl_4 by volume and an atmosphere of pure CH_3Cl. These curves show a shift toward higher voltage in going from air to CCl_4 to CH_3Cl as well as a decrease in slope. They illustrate the well-known effect of halogens in suppressing corona.

An extensive series of measurements was made first in air and later in CH_3Cl at pressures from 1 to 5 atmosphere and temperatures from room temperature to 300°C. The temperatures chosen were room temperature, 100°C, 200°C and 300°C. At each of these temperatures the volt ampere characteristics were measured at 1, 2, 3 and 4 atmospheres pressure with a small flow of gas through the tube. Figures 4 and 5 show the original observations. The curves at any one temperature show a shift towards higher voltages and a decrease in slope as the pressure is increased.

It was found that over the range of corona currents from 0.1×10^{-3} to 1×10^{-3} amp. the volt ampere characteristics gave straight lines when plotted on log-log paper. It seemed reasonable to assume that temperature and pressure had no effect on the discharge other than their effect on the density of the gas, that is the number of molecules between the electrodes. Thus the characteristics of the discharge should be uniquely determined by the mass of the gas between the electrodes.[1] Accordingly, at any given current the voltage should be a function of the gas density rather than of the temperature and pressure. To test this, values of voltage at a given current were plotted as a function of the density of the gas, taking the density at room temperature and 1 atmosphere as unity. The family of curves for air and methyl chloride are shown in Figs. 6 and 7, respectively. These figures show that at any given current the values of voltage plotted as a function of density fall on a smooth curve within the limits of the experimental error regardless of the temperature and pressure at which the measurements were made. The agreement is good except for the measurements at room temperature which depart considerably from the curve at the lower currents.

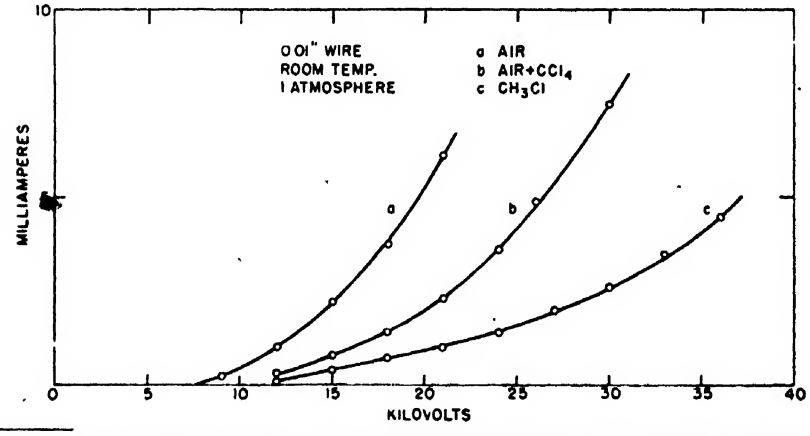


FIG. 3. Corona characteristics in various gases.

[1] Pauthenier and Rousse [Comptes Rendus 225, 1293 (1947)], show that the critical gradient for the threshold of the corona effect is a function of the specific mass of air.

All of these measurements are in a range of voltages well above the corona starting voltage. The theoretical starting voltage calculated from the equation

$$v = 33.9 g r_i \log(r_a/r_i)[1-(0.295/g r_i)],^{2}$$

where v = corona starting voltage in kv, r_a = radius of outer cylinder in cm, r_i = radius of wire in cm, g = air density taking air at stated conditions = 1 is also shown in Fig. 7.

From the curves of Figs. 6 and 7 it is possible to read off the voltages as a function of current with gas density as a parameter. When plotted on log-log paper

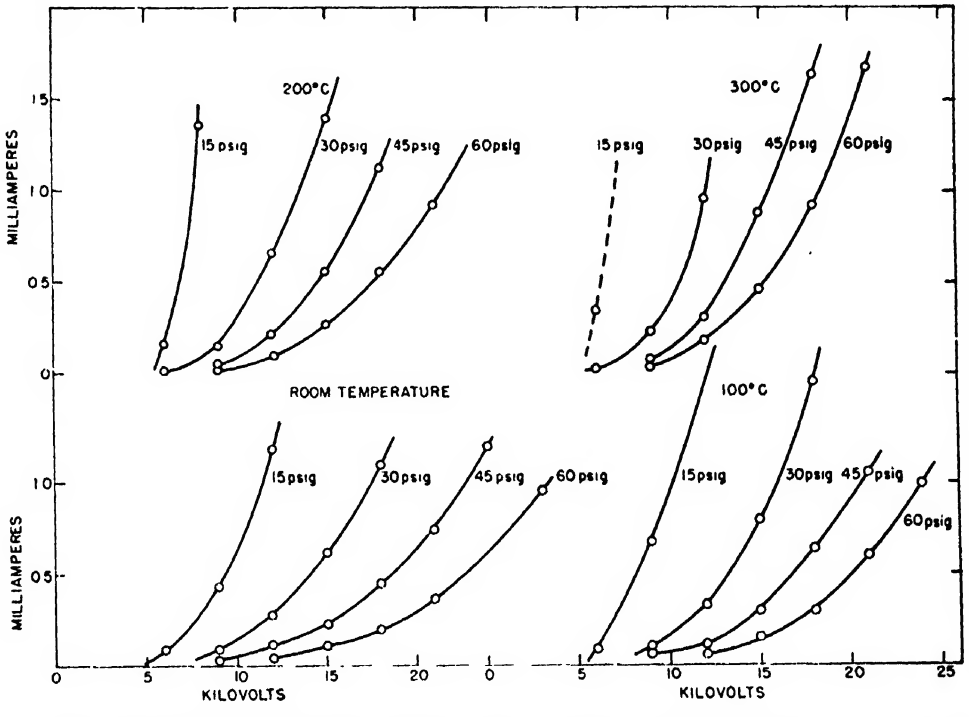


FIG. 4. Corona characteristics of 0.01" wire in air.

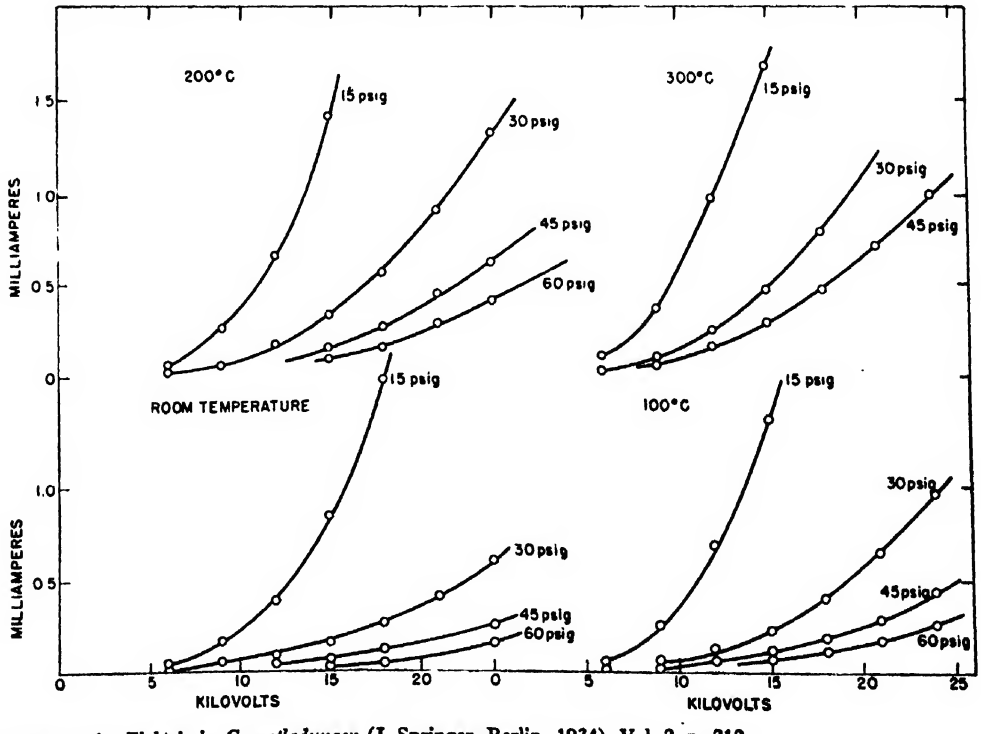


FIG. 5. Corona characteristics of 0.01" wire in methyl chloride.

[2] A. von Engel and M. Steenbeck, *Elektrische Gasentladungen* (J. Springer, Berlin, 1934), Vol. 2, p. 212.

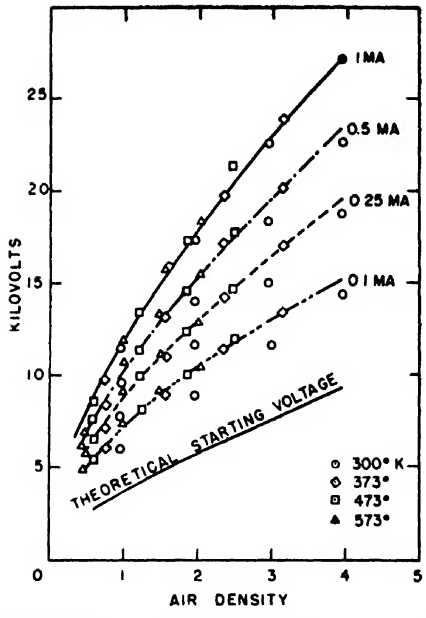


FIG. 6. Corona characteristics as a function of air density for 0.01″ wire.

TABLE I. Values of log A and K for air and methyl chloride.

d	Air	CH_3Cl	Air	CH_3Cl
	log A	log A	K	K
1	4.5	3.5	4.2	2.8
2	5.2	4.1	4.2	2.8
3	5.7	4.4	4.2	2.8
4	6.05	4.7	4.2	2.8

these volt ampere characteristics are straight lines, all having the same slope. They may be represented by an equation

$$\log i = \log A + K \log V$$

where i = current in milliamperes, V = voltage in kilovolts, A = a constant, K = a constant. They show that the current increases with the same power of the voltage at all values of density. This power is 4.2 for air and 2.8 for methyl chloride. The constants for different values of gas density are shown in Table I.

A series of measurements was carried out in air at various pressures at room temperature to determine the effect of wire size on corona characteristics. Measurements were made using tungsten wires 0.005″ and 0.02″ in diameter respectively to compare with measurements previously made with 0.01″ wire. These measurements showed that for currents of about 0.1×10^{-3} amps the effect of wire diameter on corona voltage was negligible. At 1×10^{-3} amp. the voltage increased about 3 kilovolts as the wire diameter was increased from 0.005″ to 0.02″. The gain in mechanical strength obtained by using a large diameter wire probably more than compensates for the small increase in voltage.

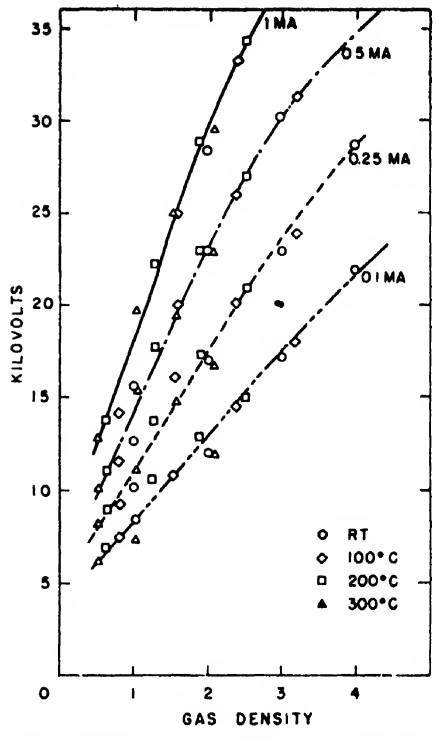


FIG. 7. Corona characteristics as a function of methyl chloride density for 0.01″ wire.

SUMMARY

The volt ampere characteristics of a corona discharge in a Cottrell precipitator are uniquely determined by the mass of gas between the electrodes and are independent of the temperature and pressure of the gas.

The corona current is a power function of the voltage. The power is independent of the gas density. Values of the power are given for air and methyl chloride.

Variations in corona characteristics with wire sizes are small at voltages above the corona starting voltage.

The Radiation from a Transverse Rectangular Slot in a Circular Cylinder*

SAMUEL SILVER AND WILLIAM K. SAUNDERS
Division of Electrical Engineering and the Antenna Laboratory, University of California, Berkeley, California
(Received January 27, 1950)

General results obtained previously for slots of arbitrary shape are applied to the case of a transverse rectangular slot in a circular cylinder. It is shown that the principal transverse plane pattern of such a slot in which the excitation has only a circumferential tangential electric field component is the same as the pattern generated by an infinite axial slot with the same circumferential excitation. Theoretical and experimental curves are given for the narrow-width half-wave-length slot.

I. INTRODUCTION

IN a previous paper[1] we developed general expressions for the external field produced by a slot of arbitrary shape in the wall of a circular cylindrical wave guide, on the assumption that the tangential electric field in the slot is a prescribed function. The present paper deals with the application of these results to the particularly simple configuration of a transverse rectangular slot: a slot that, on development of the cylinder into a plane, becomes a rectangle with boundaries respectively transverse and parallel to the axis of the cylinder. This configuration has been investigated by others by methods that differ from ours and are of less general applicability: C. H. Papas[2] has obtained expressions for the field for the case when the tangential electric field in the slot has only an axial component. G. Sinclair[3] has treated the case of a narrow rectangular slot, oriented arbitrarily on the surface of the cylinder, in which the tangential electric field is transverse to the slot axis.

The results given in this paper are restricted to the far-zone field of the slot; for this region the integrations can be performed and calculations can be made with relative ease. The computations have been made for the especially important case of a narrow slot having an axial extent of a half-wave-length. Experimental patterns have also been obtained. The data that are given here are for two values of cylinder radius that are likely to be used in microwave antenna designs.

II. THE FIELD DISTRIBUTION IN THE SLOT

The slot configuration and the various coordinates to be used are shown in Fig. 1. The slot is bounded axially by the planes $z=\pm l/2$ and in azimuth by the planes $\phi=\pm\phi_0$; or, in the notation used in I, we have for the boundaries of the slot

$$z_1=-l/2;\quad z_2=+l/2;\quad \phi_1(z)=-\phi_0;\quad \phi_2(z)=+\phi_0.$$

In formulating possible tangential electric field distributions in the slot we are guided by the consideration that at the boundaries of the slot the component of the distribution that is tangential to the boundary must be zero. In general, the excitation of the slot may be conceived as a superposition of many modes of field distribution each of which satisfies the fundamental requirement. This is analogous to the general excitation of a thin wire antenna which may be synthesized by superposition of characteristic sinusoidal distributions that satisfy the requirement that the current be zero at the ends.

The geometry of the configuration suggests that in each of the fundamental modes of excitation the field components are separable functions of ϕ and z; thus,

$$E_\phi(a,\phi,z)=f_1(\phi,z)=F_1(\phi)G_1(z), \tag{1a}$$

$$E_z(a,\phi,z)=f_2(\phi,z)=F_2(\phi)G_2(z). \tag{1b}$$

With the boundary condition at the edges of the slot in mind we are led to postulate that

$$G_1(z)=\begin{cases}\sin(p\pi z/l) & p, \text{ even integer},\\ \cos(p\pi z/l) & p, \text{ odd integer};\end{cases} \tag{2a}$$

$$F_2(\phi)=\begin{cases}\sin(p\pi\phi/2\phi_0) & p, \text{ even},\\ \cos(p\pi\phi/2\phi_0) & p, \text{ odd}.\end{cases} \tag{2b}$$

The functions $F_1(\phi)$ and $G_2(z)$ are not so clearly indicated. Certain forms are suggested for them by the point of view that the slot is a very shallow section of wave guide, in consequence of which the field configuration is like that over the cross section of a wave guide. On this basis we should have

$$F_1(\phi)=\begin{cases}\cos(q\pi\phi/2\phi_0) & q, \text{ even},\\ \sin(q\pi\phi/2\phi_0) & q, \text{ odd};\end{cases} \tag{3a}$$

$$G_2(z)=\begin{cases}\cos(q\pi z/l) & q, \text{ even},\\ \sin(q\pi z/l) & q, \text{ odd}.\end{cases} \tag{3b}$$

The limitations of the wave guide picture must be realized. It provides no basis for relating E_ϕ^{pq} to E_z^{pq} in the slot in the manner that they are related in the corresponding rectangular wave guide. Furthermore, there is some question as to whether the field components are bounded; infinities may exist at points on the slot

* Presented at the Navy Electronics Laboratory Antenna Conference, San Diego, April 1949 and at the West Coast Convention of the I.R.E. in September 1949.

[1] S. Silver and W. K. Saunders, J. App. Phys. 21, 153 (1950); this paper will be designated hereafter as I.

[2] C. H. Papas, Cruft Laboratory Technical Report No. 61, Oct. 1948 (ONR Contract N5-ori-76); J. Math. and Phys. 28, No. 4 January 1950.

[3] G. Sinclair, Ohio State University Antenna Laboratory Project Report 301-17, Sept. 1, 1949 (A.M.C. Contract W33-038 ac 16520 (17380)).

TABLE I. Values for which $H_0^{(2)\prime}(ka)/H_n^{(2)\prime}(ka) \leq 0.0001$ and $\sin(n\phi_0)/(n\phi_0) \geq 0.9$.

ka	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
n	5	6	8	9	9	10	11	12	13	13
ϕ_0 (rad.)	0.158	0.132	0.098	0.088	0.088	0.079	0.072	0.066	0.061	0.061

boundaries at which the field component is normal to the edge.[4]

III. THE FAR-ZONE FIELD

The general expressions for the far-zone field are given in the set of Eqs. (23) in I. In the present case, for a single mode of E_ϕ and E_z in the slot, the electric field components are

$$E_\theta = -\frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\sum_{n=-\infty}^{\infty}\left\{\frac{(j^{n+1}e^{-jn\phi})}{\sin\theta H_n^{(2)}(ka\sin\theta)} \times \int_{-l/2}^{l/2} G_2(\xi)e^{jk\xi\cos\theta}d\xi \int_{-\phi_0}^{\phi_0} F_2(\beta)e^{jn\beta}d\beta\right\}, \quad (4a)$$

$$E_\phi = \frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\sum_{n=-\infty}^{\infty}\left[\frac{j^n e^{-jn\phi}}{H_n^{(2)\prime}(ka\sin\theta)} \times \left\{\int_{-l/2}^{l/2} G_1(\xi)e^{jk\xi\cos\theta}d\xi \int_{-\phi_0}^{\phi_0} F_1(\beta)e^{jn\beta}d\beta + \frac{n\cot\theta}{ka\sin\theta}\int_{-l/2}^{l/2} G_2(\xi)e^{jk\xi\cos\theta}d\xi \times \int_{-\phi_0}^{\phi_0} F_2(\beta)e^{jn\beta}d\beta\right\}\right]. \quad (4b)$$

It will be observed that there is a striking difference between the fields produced by the respective modes of excitation in which there is no E_z-component in the slot and in which there is no E_ϕ-component in the slot. In the former case, when $F_2 = G_2 = 0$, the radiation field is linearly polarized having only an E_ϕ-component. In the latter case, where $F_1 = G_1 = 0$, the radiation field has both an E_ϕ- and E_z-component and, except in certain special planes, is elliptically polarized in general. These results can be correlated with the fact that the slot can be set into equivalence with a magnetic current sheet,[5] the direction of the current associated with a given tangential electric field component being transverse to that component.

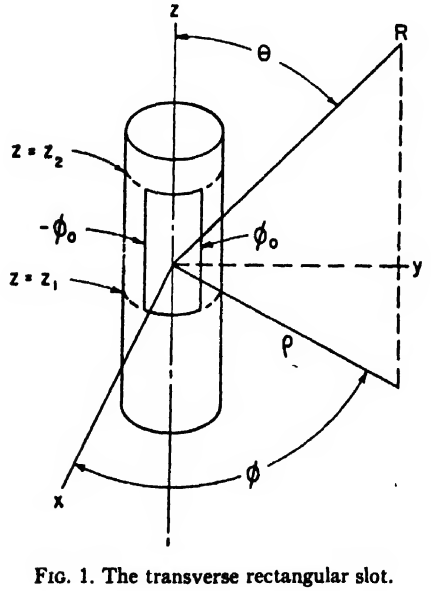


FIG. 1. The transverse rectangular slot.

There are further points of interest to be noted for the case where the tangential electric field in the slot has only an E_ϕ-component. The far-zone field is given by Eq. (4b) with F_2 and G_2 set equal to zero; on introduction of self-compensating factors of $\sin\theta$ in the appropriate places, we have

$$E_\phi = \frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\int_{-l/2}^{l/2} G_1(\xi)\sin\theta e^{jk\xi\cos\theta}d\xi \times \sum_{n=-\infty}^{\infty}\frac{(j^n e^{-jn\phi})}{\sin\theta H_n^{(2)\prime}(ka\sin\theta)}\int_{-\phi_0}^{\phi_0} F_1(\beta)e^{jn\beta}d\beta. \quad (5)$$

The integral over ξ is just the space factor of a wire antenna carrying a "magnetic current" of linear density $G_1(z)$; except for the polarization it is the same as that of an electric current having the same distribution. The summation expresses the effect of the cylinder and the transverse distribution function.

Consider now the principal transverse plane, $\theta = \pi/2$. In this plane

$$E_\phi = \left[\frac{1}{2\pi^2}\frac{e^{-jkR}}{R}\int_{-l/2}^{l/2} G_1(\xi)d\xi\right]\sum_{n=-\infty}^{\infty}\frac{j^n e^{-jn\phi}}{H_n^{(2)\prime}(ka)} \times \int_{-\phi_0}^{\phi_0} F_1(\beta)e^{jn\beta}d\beta. \quad (6)$$

The bracketed term is constant with respect to ϕ; the radiation pattern in this plane is, therefore, given by the summation alone. The latter, except for constants which do not affect the pattern, is exactly the far-zone field (in the two-dimensional sense) produced by an infinite axial slot of width $2\phi_0$ in which the excitation is constant along the slot axis, directed transverse thereto, with a transverse distribution $F_1(\phi)$. Thus we have the general result that the pattern in the plane $\theta = \pi/2$ produced by a rectangular slot having only E_ϕ-excitation is independent of the axial distribution of E_ϕ and is the same

[4] C. J. Bouwkamp, Physica XII, 467 October, 1946.

[5] S. A. Schelkunoff, Phys. Rev. 56, 308 (1939).

TABLE II. The function $\left| A \sum_{n=0}^{\infty} \frac{j^n \epsilon_n \cos n\phi}{\sin\theta H_n^{(2)\prime}(ka \sin\theta)} \right|$ for $ka = 0.8$; $A = 0.46667$ normalizes the function to unity at $\phi = 0°$, $\theta = 90°$.

θ / ϕ	10°	20°	30°	40°	50°	60°	70°	80°	90°
0°	0.5942	0.6511	0.7506	0.8506	0.9248	0.9681	0.9876	0.9970	1.0000
10	0.5938	0.6493	0.7472	0.8460	0.9199	0.9634	0.9833	0.9930	0.9961
20	0.5924	0.6440	0.7369	0.8325	0.9052	0.9492	0.9702	0.9808	0.9841
30	0.5905	0.6355	0.7203	0.8103	0.8808	0.9252	0.9478	0.9596	0.9632
40	0.5880	0.6247	0.6984	0.7802	0.8469	0.8912	0.9154	0.9286	0.9327
50	0.5853	0.6123	0.6724	0.7434	0.8044	0.8473	0.8726	0.8870	0.8914
60	0.5826	0.5994	0.6441	0.7017	0.7544	0.7941	0.8195	0.8344	0.8390
70	0.5801	0.5871	0.6157	0.6574	0.6991	0.7333	0.7569	0.7713	0.7759
80	0.5782	0.5766	0.5895	0.6139	0.6421	0.6679	0.6875	0.7000	0.7041
90	0.5770	0.5686	0.5678	0.5751	0.5879	0.6027	0.6159	0.6248	0.6280
100	0.5766	0.5638	0.5526	0.5449	0.5423	0.5443	0.5488	0.5527	0.5543
110	0.5768	0.5623	0.5449	0.5262	0.5107	0.5004	0.4952	0.4931	0.4929
120	0.5777	0.5638	0.5445	0.5203	0.4965	0.4770	0.4635	0.4561	0.4542
130	0.5790	0.5675	0.5499	0.5255	0.4992	0.4758	0.4576	0.4472	0.4442
140	0.5806	0.5725	0.5591	0.5382	0.5142	0.4917	0.4731	0.4626	0.4594
150	0.5821	0.5778	0.5694	0.5540	0.5347	0.5161	0.4999	0.4912	0.4885
160	0.5835	0.5823	0.5787	0.5686	0.5544	0.5403	0.5272	0.5207	0.5188
170	0.5843	0.5854	0.5850	0.5787	0.5682	0.5574	0.5468	0.5419	0.5407
180	0.5846	0.5865	0.5872	0.5823	0.5731	0.5635	0.5539	0.5496	0.5486

TABLE III. The function $\left| A \sum_{n=0}^{\infty} \frac{j^n \epsilon_n \cos n\phi}{\sin\theta H_n^{(2)\prime}(ka \sin\theta)} \right|$ for $ka = 2.5$; $A = 0.13630$ normalizes the function to unity at $\phi = 0°$, $\theta = 90°$.

θ / ϕ	10°	20°	30°	40°	50°	60°	70°	80°	90°
0°	0.7122	0.9167	0.9046	0.9722	0.9982	0.9852	0.9864	0.9952	1.0000
10	0.7087	0.9137	0.9027	0.9667	0.9948	0.9842	0.9846	0.9920	0.9962
20	0.6983	0.9043	0.8967	0.9505	0.9837	0.9807	0.9797	0.9835	0.9860
30	0.6814	0.8876	0.8889	0.9251	0.9624	0.9715	0.9717	0.9717	0.9722
40	0.6589	0.8626	0.8774	0.8931	0.9284	0.9518	0.9581	0.9571	0.9565
50	0.6320	0.8278	0.8621	0.8590	0.8819	0.9157	0.9325	0.9354	0.9355
60	0.6023	0.7823	0.8403	0.8271	0.8275	0.8603	0.8868	0.8970	0.8997
70	0.5720	0.7258	0.8080	0.7991	0.7750	0.7913	0.8179	0.8335	0.8389
80	0.5435	0.6595	0.7604	0.7710	0.7340	0.7239	0.7361	0.7487	0.7541
90	0.5194	0.5868	0.6943	0.7330	0.7042	0.6757	0.6655	0.6654	0.6669
100	0.5019	0.5138	0.6096	0.6741	0.6714	0.6466	0.6246	0.6120	0.6086
110	0.4922	0.4498	0.5113	0.5875	0.6151	0.6129	0.5989	0.5871	0.5828
120	0.4906	0.4060	0.4133	0.4763	0.5223	0.5453	0.5517	0.5503	0.5496
130	0.4956	0.3910	0.3389	0.3590	0.3984	0.4326	0.4539	0.4639	0.4673
140	0.5049	0.4032	0.3161	0.2776	0.2769	0.2954	0.3146	0.3269	0.3314
150	0.5155	0.4313	0.3456	0.2803	0.2341	0.2107	0.2022	0.2011	0.2012
160	0.5253	0.4616	0.3955	0.3420	0.2947	0.2602	0.2371	0.2234	0.2191
170	0.5319	0.4838	0.4358	0.4005	0.3688	0.3450	0.3298	0.3205	0.3179
180	0.5343	0.4918	0.4509	0.4230	0.3973	0.3801	0.3693	0.3627	0.3612

as that produced by an infinite axial slot having the same transverse distribution. This provides a rigorous basis for the conjecture of Sinclair,[6] Papas and King[7] and others that such a relationship exists in the case of a narrow rectangular slot in which $F_1(\phi)$ is a constant.

IV. THE NARROW RECTANGULAR SLOT

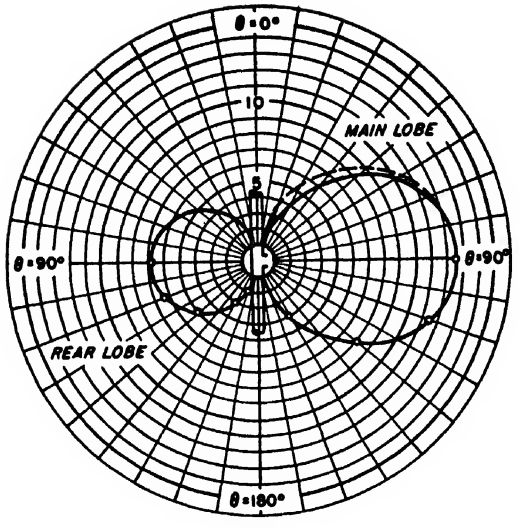
When the transverse dimension, $2a$, of the slot is small compared with the wave-length and the slot length, the significant mode of excitation is that in which there is only an E_ϕ-component that is uniform across the slot and has a sinusoidal distribution along its length. Such a slot is of particular interest because of its practical applications; the element a half-wave-length long is used extensively in microwave antennas.

It is customary to express the transverse characteristic $F_1(\phi)$ in terms of a voltage V; thus we have for the excitation

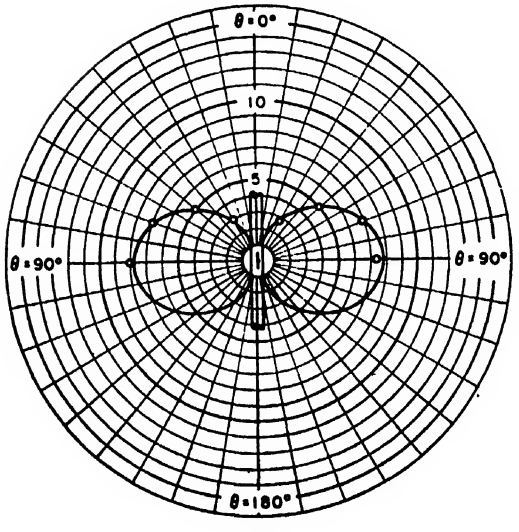
$$F_1(\phi)G_1(z) = (V/2a\phi_0)G_1(z). \tag{7}$$

On substituting into Eq. (5) we find the far-zone field to

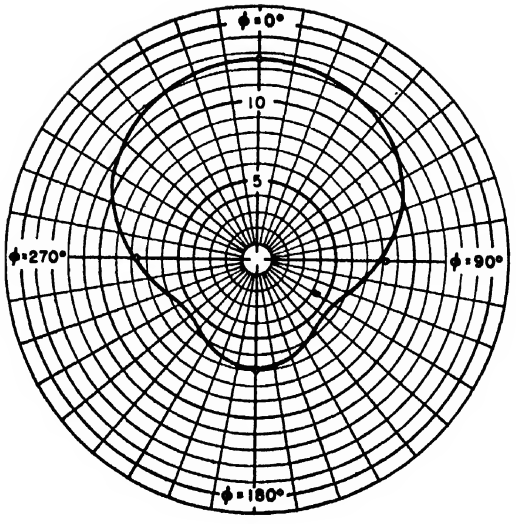
[6] G. Sinclair, Proc. I.R.E. **36**, 1487 (1948).
[7] C. H. Papas and R. W. P. King, Cruft Laboratory Technical Report No. 32, March 7, 1948.



(a)



(b)



(c)

FIG. 2. Radiation pattern ($|E|$) of the narrow half-wave slot; $ka=0.8$; (a) principal H-plane ($\phi=0, \pi$); (b) H-plane ($\phi=\pm\pi/2$); (c) principal E-plane ($\theta=\pi/2$). Solid line is the theoretical curve.

be

$$E_\phi=\frac{V}{2\pi^2 a}\frac{e^{-jkR}}{R}\int_{-l/2}^{l/2}G_1(\xi)\sin\theta e^{jk\xi\cos\theta}d\xi \times\sum_{n=-\infty}^{\infty}\frac{j^n e^{-jn\phi}}{\sin\theta H_n^{(2)\prime}(ka\sin\theta)}\frac{\sin n\phi_0}{n\phi_0}. \quad (8)$$

If ϕ_0 is sufficiently small, $\sin n\phi_0/n\phi_0\approx 1$ over the range of values of n for which the terms of the series are significant. This can be seen in Table I which shows the dependence on ka of the number of terms required for five figure accuracy and the associated value of ϕ_0 for which $\sin n\phi_0/n\phi_0\geq 0.9$ over the range of n. From a practical standpoint, therefore, we can ignore the $\sin n\phi_0/n\phi_0$ variation. Combining terms of positive and negative n, we can write finally for the radiation field of the narrow slot

$$E_\phi=\frac{V}{2\pi^2 a}\frac{e^{-jkR}}{R}\int_{-l/2}^{l/2}G_1(\xi)\sin\theta e^{jk\xi\cos\theta}d\xi \times\sum_{n=0}^{\infty}\frac{j^n\epsilon_n\cos n\phi}{\sin\theta H_n^{(2)\prime}(ka\sin\theta)};\quad \begin{cases}\epsilon_n=1, & n=0,\\ =2, & n>0.\end{cases} \quad (9)$$

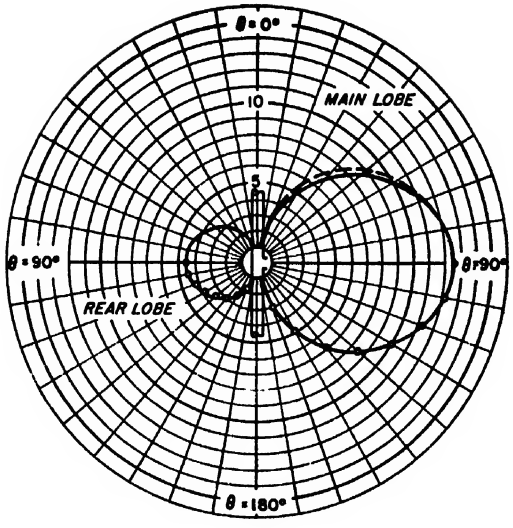
In its reduced form the series is a space factor dependent on the cylinder alone. With the series tabulated as a function of ka, θ, and ϕ the computation of the field produced by narrow axial slots becomes a simple straightforward task. Tables II and III give the function in a $10°\times 10°$ mesh in θ and ϕ for two values of ka used in our experimental work. The latter correspond to values that are likely to be used for microwave antennas.

The Half-Wave-Length Slot

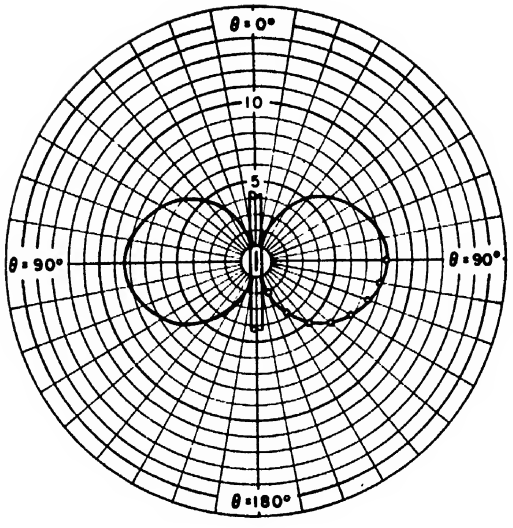
When the length of the slot is $\lambda/2$, $G_1(z)=\cos(\pi z/l)$ and the far-zone field is

$$E_\phi=\frac{V}{2\pi^2 a}\frac{e^{-jkR}}{R}\left[\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right]\times\sum_{n=0}^{\infty}\frac{j^n\epsilon_n\cos n\phi}{\sin\theta H_n^{(2)\prime}(ka\sin\theta)}. \quad (10)$$

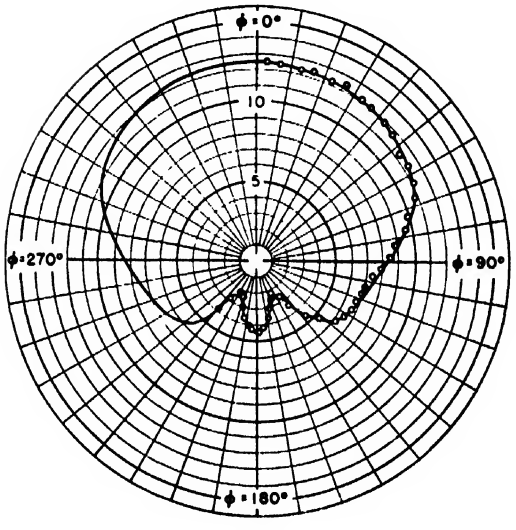
The bracketed term will be recognized to be the space factor of a half-wave-length wire antenna. The effect of the cylinder can be seen in the patterns shown in Figs. 2 and 3. Figures (2a) and (3a) are principal H-plane patterns ($\phi=0, \pi$); the dashed line is one quadrant of the pattern of the half-wave radiator in free-space. It is seen that the main lobe of the slot pattern is closer to the half-wave radiator pattern in the case of the larger cylinder than in the case of the smaller one. The corre-



(a)



(b)



(c)

FIG. 3. Radiation pattern ($|E|$) of the narrow half-wave slot; $ka=2.5$; (a) principal H-plane ($\phi=0, \pi$); (b) H-plane ($\phi=\pm\pi/2$); (c) principal E-plane ($\theta=\pi/2$).

spondence may be expected to increase with increasing cylinder radius, for the pattern of a corresponding slot in an infinite plane sheet is just that of the half-wave radiator. The principal transverse patterns ($\theta=\pi/2$) are given in Figs. (2c) and (3c).

The circles are the experimental points. The excellent agreement between the measured and theoretical values shown here was also obtained in the many other patterns that were investigated.[8] Figure 2 pertains to a probe-fed slot of width 0.006 in. cut in the wall of a $\frac{5}{16}$ in. O.D. coaxial line. The operating wave-length of 3.12 cm was chosen so that the value of ka would coincide with one of those for which transverse plane patterns were measured by Sinclair[6] in his study of the relation between the finite slot and the infinite slot. The computed pattern shown in Fig. (2c) will be found to be in excellent agreement with Sinclair's data and the pattern of an infinite slot.

The patterns in Fig. 3 were obtained for a slot of width .012 in. cut into the wall of a circular wave guide of O.D. 1 in. The circular wave guide was fed by way of a standard type of transition section from a rectangular wave guide and the slot was cut along the extended center line of the narrow side of the rectangular guide. The operating wave-length was 3.2 cm.

[8] The reader is referred to our report, "The radiation from a rectangular slot in a circular cylinder," Univ. of Calif. Antenna Laboratory Report No. 158. (Bu Ships Contract NO-bsr-39401) October 26, 1949 for a more complete set of patterns.

V. ACKNOWLEDGMENTS

We take pleasure in expressing our appreciation to Mr. J. Priedigkeit of the Antenna Laboratory for his assistance with the measurements. The work was supported in part by the Bureau of Ships under Contract No-bsr-39401 with the University.

The Thermal Conductivity of Soils

ANDREW GEMANT
Research Department, The Detroit Edison Company, Detroit, Michigan
(Received January 31, 1950)

The thermal conductivity of soils of varying moisture content is calculated on a physical basis. The results of this computation are, in a range of moisture contents from 5 to 25 percent, in agreement with experimental data in the literature. The influence of soil composition on conductivity is also explained by the theory. The use for practical purposes of an average resistivity value of 60 to 70 thermal ohms appears justified in the light of this analysis.

THE thermal conductivity of soils is an essential factor in determining the heat dissipation in various industrial installations in the ground. Such installations are, among others, underground high voltage cable lines, steam lines, and cross-country oil pipe lines.

For the design and operation of such installations a knowledge of the conductivity of the soil is necessary. The value varies with the composition of the soil, whether sand, loam, or clay, but primarily it depends on the moisture content and thus varies with the season. Pertinent data referring to various soils are presented by Nelson.[1]

For many purposes, particularly in the cable field, it is customary to use average values. An often used value for the resistivity has been 120 thermal ohms, whereas in recent years a value of 80 is more frequently used. Recent determinations by Sanderson, Sticher, and McGrath[2] indicate that still lower values, about 60 thermal ohms, might be justified.

Because of the importance of soil conductivity, it was thought of interest to calculate its value on a physical basis. Such a computation might also indicate the right order of magnitude of the average figures to be used for practical purposes.

For the following considerations both thermal conductivities λ and resistivities ρ will occur. The former is usually expressed in cal./cm sec. °C, the latter in cm °C/watt, or thermal ohms. In these units

$$\rho = 0.239/\lambda. \tag{1}$$

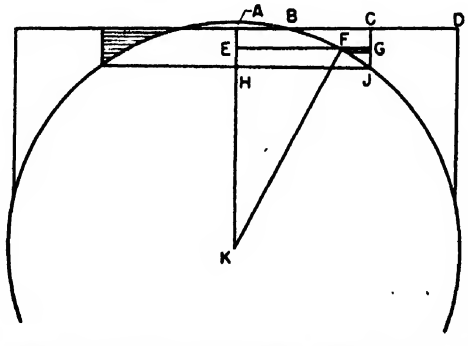


FIG. 1. Assumed geometric arrangement of solid grains and water in soil.

As mentioned, λ increases with increasing moisture content. For dry soil, λ is in the order 0.0007, increasing with increasing moisture content up to about 0.005. An apparent paradox may be mentioned at this point. One would expect that, since water increases the thermal conductivity, water itself should have a conductivity higher than soaked soil. This, however, is not the case, λ for water being 0.0014.

This apparent contradiction can be explained in the following manner. Dry soil has a low conductivity because air, a poor thermal conductor, separates the solid grains of the soil. If one adds water to the soil, hence gradually replacing the air by water, the conductivity increases. Finally, we arrive at soil saturated with water. Of the two components of such a system, the solid has a higher conductivity than the liquid; hence replacing the solid by the liquid will result in lower conductivity. The statement above, namely, that the solid component of soils is a better thermal conductor than water, is corroborated by data, as will be shown below.

The calculation of the conductivity of composite structures[3,4] can be carried out with a greater or smaller degree of rigorousness. Absolute rigorousness is impossible because of the irregularity of the shape and distribution of the grains. For the present purpose an approximate calculation was deemed sufficient.

The soil is assumed as being composed of solid spherical particles of radius $KF=r$ (Fig. 1). In reality, the grains are more likely to be oblate ellipsoids. The arrangement of the spheres is assumed such that each sphere touches six neighbors. In reality, the packing will be closer; our arrangement, however, does not deviate too much from reality, since it yields a porosity of 43 percent against about 38, found experimentally. If the spheres were to touch at a geometrical point only, the resistivity of dry soil would be close to that of air which is contrary to experience. The spheres, therefore, are assumed to be chipped and contact over the circular areas of radius $AB=a$. For the sake of simplicity the length $AK=AD$ is termed unity; the cubical volume housing the chipped sphere is then 8. The actual grain size is not involved in the following calculations.

[1] W. L. Nelson, Oil and Gas J. 44, 110 (September, 1945).
[2] Sanderson, Sticher, and McGrath, Trans. A.I.E.E. 67, 487 (1948).
[3] J. B. Austin, *Symposium on Thermal Insulating Materials* (American Society for Testing Materials, Philadelphia, 1939).
[4] W. Meissner and G. U. Schubert, "Physics of solids I," FIAT Rev. German Sci., 212 (1947).

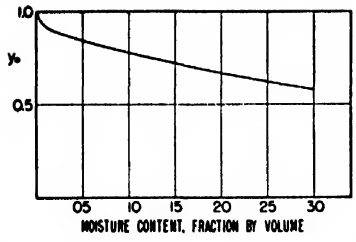


FIG. 2. Thickness of water layer between grains as a function of moisture content.

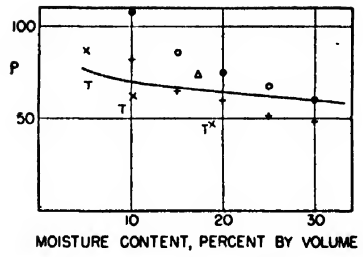


FIG. 3. Resistivity of soil *vs.* moisture content. (Solid line is calculated, points are experimental data from literature.)

If moisture is present, it will, owing to capillarity,[5] be located around the circles of radius a, occupying a wedge-shaped ring, the cross section of which is BCJ, shown also by the cross-hatched area in Fig. 1. The volume of the ring increases with increasing moisture content. If h, the moisture content, designates the ratio of water volume to total volume, then the volume of the ring BCJ equals $8h/6=1.33h$.

Any point F on the sphere is characterized by its coordinates $EF=x$ and $EK=y$. The water reaches down to point J, characterized by its coordinates $HJ=x_0$ and $KH=y_0$. Obviously, y_0 is a function of the moisture content h, and for the subsequent derivation this correlation is needed. Its calculation is straightforward, as shown.

The volume of an infinitesimally thin ring-shaped layer (Fig. 1), the cross section of which is shown at FG, is given by $\pi(x_0^2-x^2)dy$, or, because of

$$x^2+y^2=r^2, \quad \pi(y^2-y_0^2)dy.$$

The volume of the total ring is given by

$$\pi\int_{y_0}^{1}(y^2-y_0^2)dy.$$

Solving the integral, one obtains:

$$1.33h=\pi(0.33-y_0^2+0.67y_0^3). \tag{2}$$

For moisture contents over about 0.2 the formula is not strictly correct, since there will be an overlapping of the neighboring ring-shaped water pools.

Figure 2 is a graphical presentation of Eq. (2) for moisture contents up to 0.30. It may be noted that $(1-y_0)$ is a measure of the thickness of the water layer between grains. Use will be made later of this graph.

The thermal resistivity is computed by taking AK as the direction of heat flow and integrating the resistance of infinitesimally thin layers normal to the y axis. Since the flow lines are not parallel to y but diverge below y_0, this method is approximate.

Consider again the elementary layer at F. If the conductivity of the solid phase is λ_s and that of the water λ_w, the resistance of the inner circle is

$$dy/(\pi x^2\lambda_s)$$

and that of the outer ring

$$dy/[\pi(x_0^2-x^2)\lambda_w].$$

The resistance of both together is obtained by adding the conductances and taking the reciprocal of the sum. By replacing x by y, and considering that $r^2=1+a^2$, one obtains

$$dy/[\pi(f^2-g^2y^2)],$$

where

$$f^2=(1+a^2)\lambda_s-y_0^2\lambda_w \tag{3}$$

and

$$g^2=\lambda_s-\lambda_w. \tag{4}$$

The resistance of all layers between $y=1$ and y_0 is then:

$$\frac{1}{\pi}\int_{y_0}^{1}\frac{dy}{f^2-g^2y^2}=\frac{1}{2\pi fg}\log\frac{(f+g)(f-gy_0)}{(f-g)(f+gy_0)}. \tag{5}$$

Disregarding chipping at four points of the "equator" of our sphere, the resistance of all layers between $y=y_0$ and 0 is obtained as

$$\frac{1}{2\pi f'g'}\log\frac{f'+g'y_0}{f'-g'y_0}, \tag{6}$$

where

$$f'^2=(1+a^2)\lambda_s \tag{7}$$

and

$$g'^2=\lambda_s. \tag{8}$$

The resistance of all layers between $y=0$ and 1 is given by the sum of (5) and (6). Since the cross-sectional area of this volume is 4, the resistivity of the structure is four times greater. Thus

$$\frac{1}{\lambda}=\frac{2}{\pi fg}\log\frac{(f+g)(f-gy_0)}{(f-g)(f+gy_0)}+\frac{2}{\pi f'g'}\log\frac{f'+g'y_0}{f'-g'y_0}. \tag{9}$$

In the derivation above, the conductance of the air phase was neglected, as well as the conductance of the four water rings located at four points of the "equator" of our sphere.

For the numerical calculations which follow, numerical values of λ_s, λ_w, and a are required. The value of

[5] M. Devienne, J. Rech. Centr. Nat. Rech. Sci., No. 8, 238 (1949).

TABLE I. Influence of composition on conductivity of soils. Density = 1.75; moisture content = 17.5 percent.

Soil	ρ observed ohms	λ_s cal./cm sec. °C	ρ calculated ohms
Sand	46	0.014	50
Sandy loam	69	0.010	66
Silt loam	99	0.0058	101

y_0 can be read from Fig. 2 if the moisture content of the soil is known.

Although the chief constituent of soils is silica, the conductivity of pure quartz crystals is not an appropriate choice for λ_s, since the grains of soils are not composed of pure quartz crystallites. Closer to reality is probably quartzite, a compact rock composed of quartz; still better perhaps, quartz-schist, a metamorphosed rock essentially of quartz; or sandstone, a rock essentially of sand and some natural cement. The conductivity of quartzite[6] is 0.014, that of quartz-schist[7] 0.0097, and that of sandstone[7] 0.011. As an average of the last two data, $\lambda_s = 0.010$ was taken. The value of λ_w, as mentioned earlier, is 0.0014.

Concerning the radius a of the circle of contact, its value can be computed from the conductivity of dry soil, λ_d, a representative value of which is 0.00066. For this case, $1/\lambda_d$ is given by the second term of Eq. (9) with $y_0 = 1$. If, now, $a \ll 1$, the corresponding equation reduces to

$$\frac{1}{\lambda_d} = \frac{2}{\pi \lambda_s} \log \frac{4}{a^2}, \tag{10}$$

from which, with $\lambda_s = 0.010$, one obtains $a = 2 \times 10^{-5}$. The area of contact is very small compared with the grain dimension, and may be neglected in the numerical computation; in other words, unity can be substituted for $(1+a^2)$ in Eqs. (3) and (7).

Equation (9) now can be used for numerical computation of the conductivity λ of soils of varying moisture content. The result, in a range of $h = 0.05$ to 0.3, that is 5 to 30 percent by volume of moisture, is shown by the curve of Fig. 3. The ordinates are ρ in thermal ohms, calculated from λ by Eq. (1).

For the sake of comparison with the calculated curve, experimental data of Krischer[6] on two samples of soil are plotted on Fig. 3. The soils measured were loamy sand of average grain size 0.085 cm. The crosses in Fig. 3 refer to a sample of density (dry) 1.6 and 38 percent porosity, the circles to a sample of density 1.7 and 34 percent porosity. In addition, a value indicated by a triangle obtained by Sanderson *et al.*[2] is plotted.

[6] O. Krischer, Beihefte z. Gesundheits—Ing. Series I, No. 33 (1934).

[7] H. A. Nancarrow, Proc. Phys. Soc. 45, 447 (1933).

A few pertinent data taken from a recent paper by M. S. Kersten[8] are also plotted on Fig. 3. They refer to a sandy soil of density 1.7 and 37 percent porosity (X) and of density 1.8 and 34 percent porosity (T).

Above five percent moisture content, the agreement between experimental and calculated data is good; below five percent, experimental data are higher. The calculation involves rather specific assumptions as to the geometry of contact points between grains; for low moisture contents where the contacts become increasingly significant, the calculation becomes less reliable. The moisture range of greatest practical importance is between 5 and 25 percent, and for this range the calculation corroborates the experimental figures rather well. If the present calculation were carried out for oblate ellipsoids, instead of spheres, or for grains bounded partly by planes, the agreement in the low moisture content range might perhaps be improved.

The influence of the composition of the soil on conductivity, according to Kersten's observations, is such that λ increases in the sequence clay—loam—sand. Soils of constant density and moisture content must be chosen for such a comparison. Table I gives in the first column the designation of the soil and in column 2 the resistivity in ohms, as found by Kersten.[8]

According to the theory given, the influence of the composition is represented by λ_s which for a sandy loam, as discussed, is 0.010. Sand is essentially quartz for which the conductivity of quartzite, 0.014, is appropriate. Silt loam consists to a large extent of kaolin which is essentially finely divided feldspar, aluminum silicate. For feldspar,[9] $\lambda_s = 0.0058$. These values are listed in the third column of Table I. The resistivities, as calculated from Eq. (9), are given in column 4; the agreement with the observed data is satisfactory.

In practical instances the conductivity of the sequence clay—loam—sand may be reversed, clay having the highest conductivity. This is not in contradiction with the data given in Table I, which refers to the same moisture content. The reversal is caused by the varying ability of soils in retaining moisture. Thus sand, having large grains, passes water easily and has, as a rule, a low moisture content; hence, a resistivity of 80 ohms or more will result. Clay, on the other hand, has very fine pores, water is easily retained and the moisture content, as a rule, is high; the resistivity may be as low as 60 or even 50 ohms. The average soil, being a sandy loam of a water content of about 20 percent, has, according to Fig. 3, a resistivity of 60 to 70 thermal ohms.

[8] M. S. Kersten, University of Minnesota Eng. Exp. Station Bulletin No. 28 (1949).

[9] Landolt-Börnstein, *Physik-Chem. Tabellen* (Verlag. Julius Springer, Berlin, Edwards' Lithoprint, 1943), Vol. II, p. 1295.

Theory of Blind Navigation by Dynamical Measurements

J. J. GILVARRY, S. H. BROWNE,* AND I. K. WILLIAMS
The Rand Corporation, Santa Monica, California
(Received February 1, 1950)

The differential equation is considered which determines the position of a vehicle from dynamical measurements of the non-gravitational acceleration **b** made internally. Three linear approximations to the gravitational field **g**(r) of the earth, which lead to explicit solutions of this equation, are considered and their limitations are discussed. An interval-wise solution (linear continuation) for trajectories of extended range is described, which is based on such linear approximations and has definite advantages in this application. The theory is applied to the trajectory of the German A10 vehicle.

CURRENT developments in rocket and jet propulsion have made blind navigation a matter of practical importance. This paper discusses the theory of blind navigation by means of dynamical measurements (measurements of forces or accelerations) made on a proof body in a reference frame internal to a vehicle. Instrumentally, such measurements are made by an accelerometer in a reference frame provided by gyroscopic or other means. A navigation system essentially of the type considered was used in the German V2. In the following discussion, the vehicle will be idealized as a windowless box.

Under the assumption of local uniformity of the gravitational field, the information derivable from the accelerometer is limited[1] by the applicability of Einstein's equivalence principle. In fact, this limitation was recognized by the investigator[2] who built the first (consciously-designed) accelerometer.

I. THE g-CORRECTION EQUATION

The accelerometer reference frame Q will be taken non-rotating for convenience, with origin at a point Q in the box. If $\mathbf{R}$ is the radius vector of the point Q referred to an inertial reference frame O, then the acceleration $d^2\mathbf{R}/dt^2$ of this point must consist of two parts: a part $\mathbf{g}$ due to the local gravitational acceleration in space, and a part $\mathbf{b}$ due to the operation of non-gravitational forces. Thus one has

$$d^2\mathbf{R}/dt^2=\mathbf{g}+\mathbf{b}. \tag{1}$$

If $\boldsymbol{\rho}$ is the radius vector in the frame Q of the proof body and m is its mass, the equation of motion of this body in the inertial frame O is

$$d^2(\mathbf{R}+\boldsymbol{\rho})/dt^2=\mathbf{g}+\mathbf{L}/m, \tag{2}$$

where $\mathbf{L}$ is the force on the proof body linking it to the frame Q. Subtraction of (1) from (2) yields

$$d^2\boldsymbol{\rho}/dt^2-\mathbf{L}/m=-\mathbf{b} \tag{3}$$

for the equation of motion of the proof body in the observer's frame Q. If the quantity on the left-hand side in Eq. (3) is defined as the reading of the accelerometer, it follows that an accelerometer reads the negative of the acceleration of the origin of its associated reference frame due to non-gravitational forces. From the method of derivation, it is clear that Eq. (3) holds independently of variation of mass, rotation, or non-rigidity of the box.

In Eq. (1) only the term $\mathbf{b}$ is susceptible to physical measurement in the box, and the observer can integrate the differential equation to yield his position vector $\mathbf{R}(t)$ only if he knows the functional form of $\mathbf{g}$ in its dependence on position coordinates and also the initial position and velocity of the box. When the term $\mathbf{g}$ is replaced by the function specifying its dependence on position coordinates in space, and the term $\mathbf{b}$ is considered a function $\mathbf{b}(t)$ of time, Eq. (1) will be referred to as the g-correction equation.

For motion of the box in the earth's neighborhood, the inertial frame O can be taken as approximately geocentric. It is convenient to select an initial point O_0 with radius vector $\mathbf{R}_0$ which, in general, is outside the earth's surface and on the curve of motion C (Fig. 1). The gravitational field corresponding to a spherical earth is

$$\mathbf{g}=-g_0(R_0^2/R^3)\mathbf{R}, \tag{4}$$

if g_0 is the value of g at O_0. The g-correction equation then becomes

$$d^2\mathbf{R}/dt^2+g_0(R_0^2/R^3)\mathbf{R}=\mathbf{b}(t). \tag{5}$$

The sphere S_0 (Fig. 1) through O_0 and concentric with the earth will be referred to as the reference sphere. For an initial point on the actual (approximately spheroidal) earth, the radius and surface gravitation of the reference sphere can be taken as the mean radius R_e and mean gravitational acceleration g_e over the earth's surface, which are defined[3] in first approximation by

$$R_e=(2/3)R_{eq}+(1/3)R_{pol}, \tag{6a}$$

$$g_e=(2/3)g_{eq}+(1/3)g_{pol}, \tag{6b}$$

where R_{eq} and R_{pol} are the equatorial and polar radii

* Now at North American Aviation, Inc., Los Angeles, California.

[1] J. J. Gilvarry, Phys. Rev. 73, 1409 (1948). A misprint exists in this letter; in the first line of the second column of p. 1410, delete "vertically."

[2] F. W. Lanchester, Proc. Phys. Soc. 19, 691 (1905).

[3] G. G. Stokes, Camb. and Dubl. Math. J. 4, 194 (1849); Trans. Camb. Phil. Soc. 8, 672 (1849); also *Mathematical and Physical Papers* (Cambridge University Press, Cambridge, 1883), Vol. II, p. 104.

respectively of the earth, and g_{eq} and g_{pol} are the corresponding gravitational accelerations. The radius R_e is, in first approximation, equal to the radius of the earth's sphere of equal volume, and defines an equivalent sphere S_e (Fig. 1) with surface gravity g_e for the earth.[4] Equation (5) is valid in a geocentric frame O' fixed in the rotating earth if the centrifugal and Coriolis accelerations (and the latitude-dependent component of $\mathbf{g}$) can be neglected.[5]

Let a Cartesian frame O_0 (Fig. 1) be defined with origin at the initial point of motion, such that the z-axis is vertical, the x-axis is horizontal in the plane determined by the z-axis and the initial velocity vector $\mathbf{v}_0$, and the y-axis is selected to make the frame right-handed. It will be assumed for simplicity in what follows that the initial velocity $\mathbf{v}_0$ is non-vanishing unless otherwise stated (and hence the convention above defines the direction of the x-axis except when the initial velocity is in the vertical direction). If $\mathbf{r}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$ is the radius vector of the box in the frame O_0, Eq. (5) becomes

$$d^2\mathbf{r}/dt^2+\omega_0^2(R_0/R)^3(R_0\mathbf{k}+\mathbf{r})=\mathbf{b}(t), \tag{7}$$

where the parameter ω_0 is defined by

$$\omega_0=(g_0/R_0)^{\frac{1}{2}}, \tag{8}$$

and

$$R=[x^2+y^2+(z+R_0)^2]^{\frac{1}{2}}, \tag{9}$$

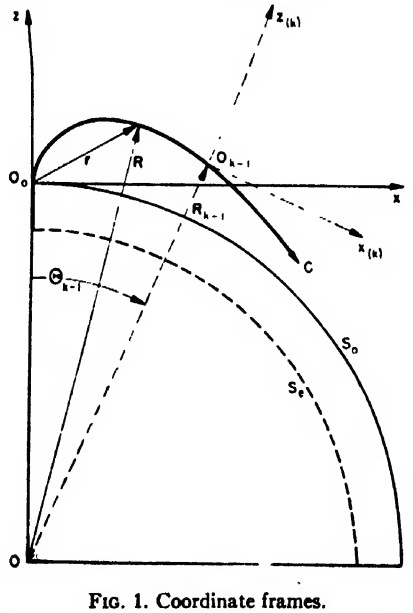


FIG. 1. Coordinate frames.

[4] The equality of volumes follows on the assumption that squares and higher powers of the ellipticity $e=(R_{eq}-R_{pol})/R_{eq}$ are negligible compared to unity. If the density of the earth's crust is assumed constant, R_e is likewise the radius of the earth's sphere of equivalent mass. Hence, Eq. (4) with $R_0=R_e$ and $g_0=g_e$ yields the correct asymptotic value GM/R^2 (G=gravitational constant, M=mass of the earth) of g at large distances from the earth.

[5] Equation (5) likewise ignores a slight perturbation due to lunar and solar attraction (similar in origin to the perturbations producing tides). The corresponding effect for artillery shells is discussed in F. R. Moulton, *New Methods in Exterior Ballistics* (University of Chicago Press, Chicago, 1926), p. 19.

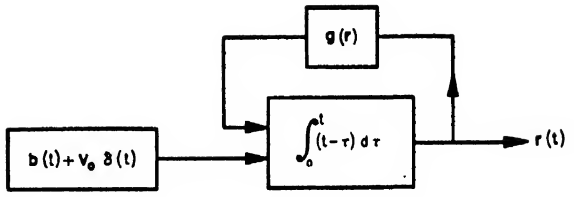


FIG. 2. Computer diagram.

with $\mathbf{b}=b_x\mathbf{i}+b_y\mathbf{j}+b_z\mathbf{k}$. The initial conditions are $\mathbf{r}_0=0$ and $\mathbf{v}_0=v_{x,0}\mathbf{i}+v_{z,0}\mathbf{k}$.

Equation (7) can be formulated as an integral equation

$$\mathbf{r}=\mathbf{v}_0t+\int_0^t(t-\tau)[\mathbf{b}(\tau)+\mathbf{g}(\mathbf{r})]d\tau, \tag{10}$$

in which $\mathbf{g}(\mathbf{r})$ is $\mathbf{g}$ of (4) with its dependence on components of $\mathbf{r}$ represented symbolically, and the cubature operator $\int_0^t(t-\tau)d\tau$ performs the operation of double integration with respect to time (as integration by parts verifies). From the standpoint of an automatic computer in the box, the sequence of operations shown diagrammatically in Fig. 2 yields the position vector $\mathbf{r}(t)$ from $\mathbf{b}(t)$ (the initial velocity $\mathbf{v}_0$ has been put in symbolically by means of the Dirac δ-function, $\delta(t)$, which is singular at $t=0$).

From the standpoint of determining $\mathbf{r}(t)$ as a closed solution in the time, Eq. (7) poses serious difficulties since $\mathbf{b}(t)$ is an arbitrary function of time and the equation is non-linear in $\mathbf{r}$. In the general case, integrals of energy and angular momentum cannot be found except as purely formal expressions Hence, the classical method of using integrals of the motion to reduce the order of the component differential equations is not possible in general. Exact solutions in some particular cases are available; if $\mathbf{b}$ is zero, for example, Eq. (7) reduces to the differential equation of an elliptical free-flight trajectory. In practical cases, however, $\mathbf{b}$ and $\mathbf{g}$ are of comparable magnitude over the trajectory as a whole (see Fig. 4), so that perturbation methods are not generally useful. Resort can be made to various numerical and iterative methods. A difference method due to Hartree,[6] which avoids intermediate computation of the velocities, can be mentioned. The integral equation, (10), can be identified in coordinate formulation with a system of non-linear Volterra integral equations of the second kind[7] to yield a general solution by iteration. Although iterative solutions have obvious limitations, a solution of this type has been applied[8] as a computational method for specific $\mathbf{b}(t)$.

In the practical problem of computer design, the non-linear $\mathbf{g}$-term of Eq. (7) presents design complications. Furthermore, a suitable approximation for $\mathbf{g}$ may be sufficient on trajectories of restricted range, or when

[6] D. R. Hartree, Mem. Manch. Lit. Phil. Soc. 77, 91 (1932).

[7] V. Volterra, *Leçons sur les Equations Intégrales* (Gauthier-Villars, Paris, 1913), p. 90.

[8] E. T. Benedikt, Phys. Rev. 74, 1213(A) (1948).

the motion is constrained to a particular surface (see Section 2). The simplest such approximation consists in replacing $\mathbf{g}$ by its constant value $-g_0\mathbf{k}$ at O_0, which yields the zero-order solution

$$\mathbf{r}=\mathbf{v}_0t-(g_0t^2/2)\mathbf{k}+\int_0^t(t-\tau)\mathbf{b}(\tau)d\tau, \qquad (11)$$

valid in a restricted neighborhood of the origin. The remainder of this paper will consider the problem of replacing $\mathbf{g}$ by linear approximations. These linear approximations will be made the basis of an interval-wise solution having certain advantages.

II. THE SPHERICAL SOLUTION

The g-correction equation, (7), has a simple and exact solution when the vehicle is constrained to move on the surface of a sphere concentric with the earth, and the corresponding solution will be referred to as a spherical solution. The sphere of motion will be taken as the reference sphere of radius R_0, on which the outward-drawn unit normal is $\mathbf{R}/R_0$ or $\mathbf{k}+\mathbf{r}/R_0$. The condition on $\mathbf{b}$ for motion on S_0 is

$$\mathbf{b}\cdot(R_0\mathbf{k}+\mathbf{r})=g_0R_0^2/R-v^2, \qquad (12)$$

which is necessary and sufficient if the initial velocity vector is tangent to S_0 at $O_0(\mathbf{v}_0=v_0\mathbf{i})$. Physically, Eq. (12) states (when divided by R) that the radial component of $\mathbf{b}$ equals the gravitational acceleration less the centrifugal acceleration on a sphere of radius R. In practice, the condition (12) can be met by means of an altimeter (of atmospheric-pressure or radar type) used to control independently the altitude of the vehicle relative to the earth. The field $\mathbf{g}_s$ on the surface of S_0 is

$$\mathbf{g}_s=-g_0\mathbf{k}-\omega_0^2\mathbf{r}, \qquad (13)$$

which will be referred to as the spherical field.

The g-correction equation in the spherical case is

$$d^2\mathbf{r}/dt^2+\omega_0^2\mathbf{r}=\mathbf{b}-g_0\mathbf{k}. \qquad (14)$$

The solution of Eq. (14) subject to the condition (12) on $\mathbf{b}$ is

$$\mathbf{r}=[(\sin\omega_0t)/\omega_0]\mathbf{v}_0-[2R_0\sin^2\omega_0t/2]\mathbf{k}+(1/\omega_0)\int_0^t\mathbf{b}(\tau)\sin\omega_0(t-\tau)d\tau, \qquad (15)$$

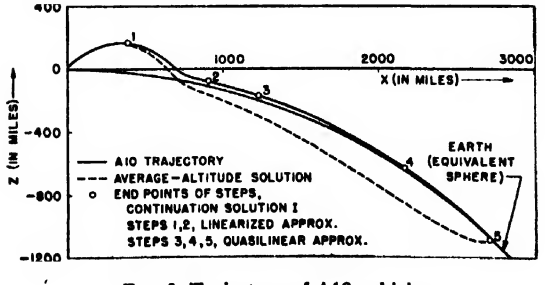


FIG. 3. Trajectory of A10 vehicle.

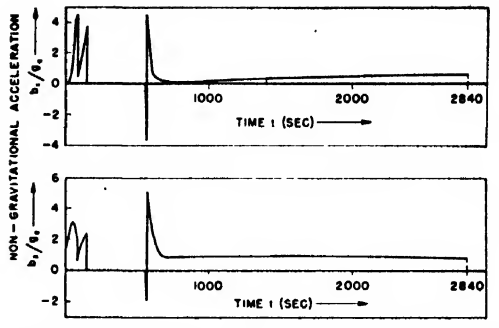


FIG. 4. Non-gravitational accelerations b_x and b_z on A10 vehicle in terms of g_e (surface gravity on equivalent sphere).

with the restriction that $\mathbf{v}_0=v_0\mathbf{i}$. If the sine functions appearing in the spherical solution are replaced by the leading terms of their Taylor expansions, the spherical solution (15) reduces to the zero-order solution (11).

As a trivial example of the spherical solution, consider a box moving on the surface of the reference sphere S_0 and such that $\mathbf{b}=0$. From Eq. (12), the speed $v_{s,0}$ is a constant

$$v_{s,0}=(g_0R_0)^{\frac{1}{2}}. \qquad (16)$$

Equation (15) in this case corresponds to a circle of radius R_0 described with the constant angular velocity ω_0 and the period

$$T_0=2\pi(R_0/g_0)^{\frac{1}{2}}, \qquad (17)$$

which is approximately 84.4 min. for an origin on the equivalent sphere S_e. Hence, this special case corresponds physically to the motion of a satellite around the earth in a circular orbit, and the speed $v_{s,0}$ will be referred to as the satellite (or circular) speed corresponding to the radius R_0.

When the constraining condition (12) is not satisfied exactly at all times, the altitude,

$$h=R-R_0, \qquad (18)$$

of the vehicle relative to the reference sphere S_0 is nonvanishing, and the spherical solution (15) yields a corresponding error. If $\delta_s\mathbf{r}$ is the true solution of Eq. (7) less the spherical solution of (15) for a given $\mathbf{b}$, then one has

$$d^2\delta_s\mathbf{r}/dt^2+\omega_0^2\delta_s\mathbf{r}=g_0\xi(\mathbf{k}+\mathbf{r}/R_0), \qquad (19)$$

where the error parameter ξ is defined by

$$\xi=1-(R_0/R)^3\sim 3h/R_0, \qquad (20)$$

and the approximation indicated is valid for $h\ll R_0$. The quantity $\xi(\mathbf{k}+\mathbf{r}/R_0)$ is the gradient of a harmonic function, and hence its magnitude assumes a maximum value on the boundary of any closed region of space (not including the center of the earth) considered. Thus, if a uniform bound H on the absolute value of h is known, then

$$M=(1-H/R_0)^{-2}-(1-H/R_0)\sim 3H/R_0 \qquad (21)$$

is a uniform bound on $|\xi(\mathbf{k}+\mathbf{r}/R_0)|$ for $H<R_0$, and hence one can show that

$$|\delta_s\mathbf{r}| \leqslant 2g_0M\sin^2\omega_0t/2, \quad (t\leqslant T_0/2), \tag{22}$$

if no error exists in the knowledge of initial conditions (the direction of $\mathbf{v}_0$ is now unrestricted). Since $|\delta_s\mathbf{r}|$ vanishes as $M\to 0$ (or $H\to 0$), the spherical solution (15) is a valid approximation in the large (or "on the average") for motion which is closely spherical, within the limitations of (22) or a closer error bound.

Over a trajectory of extended range, the spherical solution is not useful if significant systematic variations of altitude exist. The difficulty usually cannot be avoided by taking an average R corresponding to an average altitude in the $\mathbf{g}$-term of Eq. (7); this point can be illustrated by an example. Figure 3 shows the trajectory (solid curve) of the German A10 vehicle,[9,10] which is extreme in its altitude variation (185 mi.), and Fig. 4 shows the non-gravitational accelerations[11] b_x and b_z for this vehicle as a function of time. For this trajectory, a solution of the g-correction equation is shown in Fig. 3 (dashed curve) which corresponds to an average altitude of 16.9 mi. (i.e., R was taken as $R_e+16.9$ mi. in the $\mathbf{g}$-term of Eq. (7)). This particular average-altitude solution yields a negligible error at the terminal point of the motion but defines a trajectory passing under the earth's surface.

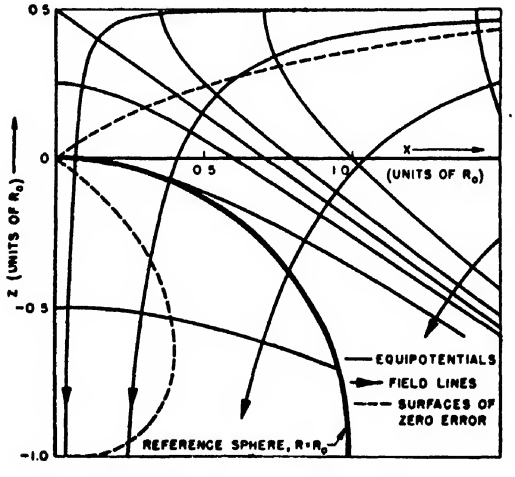


FIG. 5. Linearized field diagram.

[9] W. G. A. Perring, J. Roy. Aero. Soc. 50, 483 (1946).

[10] J. M. J. Kooy and J. W. H. Uytenbogaart, *Ballistics of the Future* (McGraw-Hill Book Co., Inc., New York, 1946), p. 399.

[11] Only a limited amount of information is available on the A10 (this vehicle was proposed but never built, and was to consist essentially of a winged V2 equipped with a booster rocket). The trajectory of Fig. 3 fits the A10 as closely as the available information permits. The quantities b_x and b_z were determined by differentiation from the assumed trajectory. Of the peaks in each curve of Fig. 4, the first two correspond to release of the booster and power cut-off in that order, and the ones at 560 seconds correspond to re-entry into appreciable atmosphere. Rotation of the earth and latitude variation of g have been completely neglected in the data of Figs. 3 and 4.

III. THE LINEARIZED APPROXIMATION

It is clear from the preceding discussion that the essential problem in connection with the g-correction equation exists when significant variations in altitude above the earth occur. This section discusses an approximate analytic solution which is superior to the zero-order approximation (11) when variations in altitude above the earth occur.

The field $\mathbf{g}(\mathbf{r})$ of (4) can be expanded about the origin O_0 to yield

$$\mathbf{g}(\mathbf{r}) = -g_0\mathbf{k}-\omega_0^2[\mathbf{r}-3(\mathbf{k}\cdot\mathbf{r})\mathbf{k}]-(3\omega_0^2/2R_0) \times\{[5(\mathbf{k}\cdot\mathbf{r})^2-\mathbf{r}^2]\mathbf{k}-2(\mathbf{k}\cdot\mathbf{r})\mathbf{r}\}+\cdots. \tag{23}$$

If only linear terms are retained in the Taylor expansion (23), the corresponding field $\mathbf{g}_L$, defined by

$$\mathbf{g}_L(\mathbf{r}) = -g_0\mathbf{k}-\omega_0^2[x\mathbf{i}+y\mathbf{j}-2z\mathbf{k}], \tag{24}$$

will be referred to as the linearized field. Under this approximation for $\mathbf{g}$, the coordinate formulation of the g-correction equation becomes

$$\ddot{x}+\omega_0^2x = b_x, \tag{25a}$$

$$\ddot{z}-2\omega_0^2z = b_z-g_0, \tag{25b}$$

with the y-equation symmetric to that in x. These differential equations, which are valid in some neighborhood of the origin, will be referred to as the linearized equations.

The linearized solution will be taken as

$$x = (v_{x,0}/\omega_0)\sin\omega_0t+(1/\omega_0) \times\int_0^t b_x(\tau)\sin\omega_0(t-\tau)d\tau, \tag{26a}$$

$$y = (1/\omega_0)\int_0^t b_y(\tau)\sin\omega_0(t-\tau)d\tau, \tag{26b}$$

$$z = (v_{z,0}/2^{\frac{1}{2}}\omega_0)\sinh 2^{\frac{1}{2}}\omega_0t-R_0\sinh^2\omega_0t/2^{\frac{1}{2}} +(1/2^{\frac{1}{2}}\omega_0)\int_0^t b_z(\tau)\sinh 2^{\frac{1}{2}}\omega_0(t-\tau)d\tau. \tag{26c}$$

The case of a flat earth corresponds to $R_0\to\infty$, or equivalently $\omega_0\to 0$. Equations (26) reduce directly to the zero-order solution (11) for a flat earth as $\omega_0\to 0$ if note is taken of the limits $(\sin u)/u\to 1$, $(\sinh u)/u\to 1$ as $u\to 0$.

The linearized field $\mathbf{g}_L$ is symmetric about the z-axis. In the meridian x, z plane, the equipotentials of the linearized field are defined by

$$(z-R_0/2)^2-x^2/2 = \text{const.}, \tag{27}$$

and the field lines are given by

$$x^2(z-R_0/2) = \text{const.} \tag{28}$$

Hence, the equipotentials form in the x, z plane a family of concentric hyperbolas and their conjugates

asymptotic to the lines $x^2=2(z-R_0/2)^2$ through the center of symmetry $(0, R_0/2)$ of the field. The linearized field diagram is shown in Fig. 5 (solid curves, field lines indicated by arrows). On the horizontal plane $z=R_0/2$, the z-component of the linearized field vanishes and above this plane the field becomes repulsive. The field diagram of Fig. 5 yields a qualitative index of the extent of the neighborhood of the origin in which the linearized equations provide a practical approximation.

In discussing the degree of approximation of the linearized solution, the error vector $\delta_L\mathbf{r}$ with components $\delta_L x$, $\delta_L y$, $\delta_L z$ will represent the true less the linearized solution. By subtracting the linearized equations, (25), from the corresponding component equations of (7), one obtains (omitting the equation for $\delta_L y$ symmetric to that in $\delta_L x$)

$$d^2\delta_L x/dt^2+\omega_0^2\delta_L x=\omega_0^2\xi x, \tag{29a}$$

$$d^2\delta_L z/dt^2-2\omega_0^2\delta_L z=\omega_0^2\zeta R_0, \tag{29b}$$

where ξ is given by (20) and the error parameter ζ is defined by

$$\zeta=1-2z/R_0-(1+z/R_0)(R_0/R)^3\sim 3(r^2-3z^2)/2R_0^2. \tag{30}$$

Physically, the right-hand sides of Eqs. (29a) and (29b) are the accelerations neglected by $\mathbf{g}_L$ on the x- and z-axes respectively. If bounds $\max(|\xi x|)$ and $\max(|\zeta R_0|)$ are known from sources other than the linearized solution itself, then one has

$$|\delta_L x|\leqslant 2\max(|\xi x|)\sin^2\omega_0 t/2, \quad (t\leqslant T_0/2); \tag{31a}$$

$$|\delta_L z|\leqslant \max(|\zeta R_0|)\sinh^2\omega_0 t/2^{\frac{1}{2}}. \tag{31b}$$

Both ξx and ζR_0 are harmonic functions, so their maximum and minimum values must be on the boundary of any closed region of space (not containing the center of the earth) considered.

It follows from (29) that the linearized x- and z-solutions are exact for motion of the box on surfaces such that $\xi x=0$ and $\zeta=0$ respectively. If one disregards the trivial locus represented by the z-axis, the reference sphere $R=R_0$ is (from Eq. (20)) the surface of zero error $\xi=0$ for the linearized x-solution (and for the linearized y-solution, by symmetry). The surface defined by $\zeta=0$ is a surface of revolution about the z-axis, and consists of two sheets intersecting the origin O_0 which are shown (dashed curves) in x, z coordinates on the linearized field diagram of Fig. 5. The upper open branch in Fig. 5 is asymptotic to the line $z=R_0/2$, and the upper and lower branches behave in the neighborhood of the origin O_0 as segments (for x positive) of the two straight lines $x^2=2z^2$. It is seen that no surface exists on which all components of the linearized solution are exact.

The leading term of the Taylor expansion of the error $\delta_L\mathbf{r}$ is given by

$$\delta_L\mathbf{r}=R_0[2v_{x,0}v_{z,0}\mathbf{i}+(v_{z,0}^2-2v_{x,0}^2)\mathbf{k}]\times(\omega_0 t)^4/4v_{e,0}^2+\cdots, \tag{32}$$

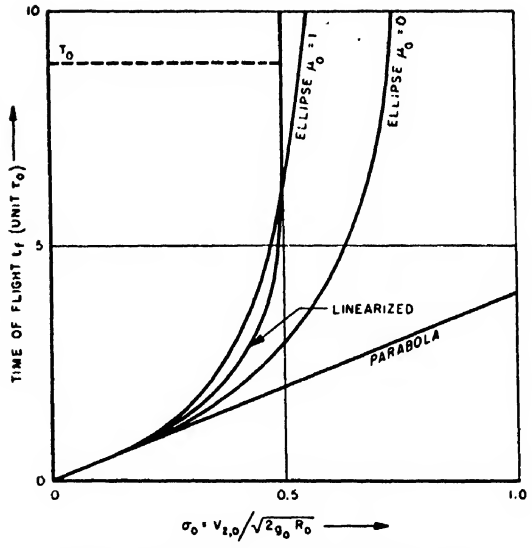


FIG. 6. Times of flight to horizontal plane $z=0$ (ballistic case, $\mathbf{b}=0$).

where $v_{e,0}$ is the escape (or parabolic) speed

$$v_{e,0}=(2g_0R_0)^{\frac{1}{2}}. \tag{33}$$

The leading terms of $\delta_L x$ and $\delta_L z$ vanish for motions initially tangent to the corresponding surfaces of zero error, respectively, and hence never vanish simultaneously. For $v_0\neq 0$ the absolute error $|\delta_L\mathbf{r}|=O(t^4)$ is never of order higher than t^4, and thus the Taylor expansion of the linearized solution agrees with the true Taylor expansion of $\mathbf{r}$ through terms of order t^3, as compared to the zero-order solution (11), which agrees through terms of order t^2. Since the arc length s on the curve of motion satisfies $s=O(t)$ when $v_0\neq 0$, the linearized trajectory has contact of order not less than the third with the actual trajectory at the origin, and hence the linearized solution is essentially an osculating approximation.

For the case $\mathbf{b}=0$ (ballistic trajectory), the linearized trajectory can be compared with the exact solution representing an ellipse in space and with the corresponding parabola. It is convenient to introduce the dimensionless parameters

$$\mu_0=v_{x,0}/(g_0R_0)^{\frac{1}{2}}, \quad \sigma_0=v_{z,0}/(2g_0R_0)^{\frac{1}{2}}, \tag{34}$$

for the initial velocities, where the denominators in μ_0 and σ_0 are, respectively, the satellite speed $v_{s,0}$ of (16) and the escape speed $v_{e,0}$ of (33) (one has $\sigma_0^2+\mu_0^2/2\leqslant 1$ if $v_0\leqslant v_{e,0}$). The linearized solution for $\mathbf{b}=0$ can be written

$$x=R_0\mu_0\sin\omega_0 t, \tag{35a}$$

$$z=(R_0/2)[1-(1-4\sigma_0^2)^{\frac{1}{2}}\cosh(t/\tau_0-\tanh^{-1}2\sigma_0)], \tag{35b}$$

in which the parameter τ_0 is

$$\tau_0=1/2^{\frac{1}{2}}\omega_0, \tag{36}$$

and it is assumed that $|\sigma_0|<1/2$. If $\sigma_0=1/2$, the z-solu-

tion becomes

$$z=(R_0/2)[1-e^{-t/\tau_0}], \tag{37}$$

and hence the trajectory defined by the linearized solution approaches the plane $z=R_0/2$ asymptotically as $t\to\infty$. The time constant τ_0 for z to reach the fraction $1-1/e$ of its asymptotic value in this case is approximately 9.5 min. for an initial point O_0 on the surface of the equivalent sphere S_e. For $\sigma_0>1/2$, the linearized trajectory crosses the plane $z=R_0/2$ and is unbounded in z; thus the linearized approximation implies an escape speed $v_{e,0}/2$ which is one-half the true value (33) on an inverse-square field. For $0<\sigma_0<1/2$, the time of flight t_f corresponding to intersection with the x,y plane,

$$t_f=2\tau_0 \tanh^{-1}2\sigma_0, \tag{38}$$

is compared in Fig. 6 with the corresponding time of flight $4\sigma_0\tau_0$ for a parabola, and, for a few values of μ_0, with the corresponding time of flight on an inverse-square field.

In Fig. 7, linearized trajectories from Eqs. (35) are shown (solid curves) in two families ($\sigma_0=0.3$ and $\mu_0=0.2$). From Eq. (35a) one notes that all the linearized trajectories for $\mathbf{b}=0$ intersect the z-axis at time $t=T_0/2$. For σ_0 small and thus t_f small relative to $T_0/2$ (as in the family $\sigma_0=0.3$), this intersection point is far below the x,y plane, and above this plane, the linearized trajectories show a fairly close resemblance in the large to the corresponding ellipses (dashed curves). The intersection with the z-axis is at the origin O_0 for $t_f=T_0/2$, which fixes a critical value $(\tanh\pi/2^{1/2})/2=0.488$ of σ_0 at which the upward and downward branches of the linearized trajectory above the x,y plane coincide. It is clear that for σ_0 close to $1/2$, the linearized trajectories for $\mathbf{b}=0$ are a very poor approximation in the large to the corresponding ellipses, although they are superior to the corresponding parabolas (dot-and-dashed curves) in a neighborhood of the origin.

IV. THE QUASILINEAR APPROXIMATION

The linearized approximation is based on the assumption that the gravitational acceleration $\mathbf{g}$ is a slowly varying function of position. This section discusses a linear approximation based on the additional assumptions that the radius of curvature of the trajectory is large and is a slowly varying function. The approximation will be referred to as the quasilinear approximation. For simplicity, only motion confined to the vertical x,z plane will be considered.

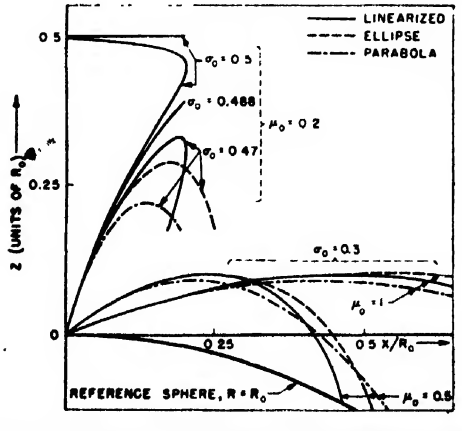


FIG. 7. Comparison of linearized trajectories with ellipses and parabolas (ballistic case, $\mathbf{b}=0$).

If the curve of motion is an analytic curve at the origin O_0, the radius vector $\mathbf{r}$ to a point on the curve can be expanded in the series[12]

$$\mathbf{r}=\alpha_0 s+(\beta_0/2\rho_0)s^2+\cdots, \tag{39}$$

where s is the arc length on the curve, α_0 and β_0 are the unit tangent and unit principal normal at O_0 respectively, and ρ_0 is the radius of curvature at O_0. The terms of order s^3 and higher in (39) involve first and higher derivatives with respect to s and powers above the first of the curvature $1/\rho$. If the series (39) for $\mathbf{r}$ is substituted in (23) for $\mathbf{g}(\mathbf{r})$, one obtains

$$\mathbf{g}(\mathbf{r})=-g_0\mathbf{k}-\omega_0^2[\alpha_0-3(\alpha_0\cdot\mathbf{k})\mathbf{k}]s - \frac{3\omega_0^2}{2}\left\{\frac{\beta_0/3-(\beta_0\cdot\mathbf{k})\mathbf{k}}{\rho_0} + \frac{[5(\alpha_0\cdot\mathbf{k})^2-1]\mathbf{k}-2(\alpha_0\cdot\mathbf{k})\alpha_0}{R_0}\right\}s^2+\cdots, \tag{40}$$

where the terms given explicitly represent $\mathbf{g}$ under the assumption that terms of order s^3 or higher are negligible. Consider the canonical coordinates x_1 and x_2 (defined as coordinates in the direction of α_0 and β_0 respectively) of the curve at the origin. The coordinates x_1 and x_2 are given in terms of x and z by

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}=\begin{bmatrix} \alpha_{x,0} & \alpha_{z,0} \\ \beta_{x,0} & \beta_{z,0} \end{bmatrix}\begin{Bmatrix} x \\ z \end{Bmatrix}, \tag{41}$$

where the elements of the direction-cosine matrix are components of α_0 and β_0; these coordinates are likewise given by

$$x_1=s, \quad x_2=s^2/2\rho_0, \tag{42}$$

from (39) under the approximation that terms of order s^3 or higher are negligible. By comparison of (41) and (42), one has the approximation

$$s=\alpha_{x,0}x+\alpha_{z,0}z, \tag{43a}$$

$$s^2=2\rho_0[\beta_{x,0}x+\beta_{z,0}z], \tag{43b}$$

which expresses both s and s^2 linearly in terms of x and z. Under the substitution (43), the terms given explicitly in (40) yield the quasilinear field $\mathbf{g}_Q$ defined (in matrix notation) by

$$\mathbf{g}_Q=-\begin{Bmatrix} 0 \\ g_0 \end{Bmatrix}-\omega_0^2\mathbf{A}_0\begin{Bmatrix} x \\ z \end{Bmatrix} \tag{44}$$

[12] W. C. Graustein, *Differential Geometry* (Macmillan Company, Inc., New York, 1947), p. 39.

where the coefficient matrix $\mathbf{A}_0$ is

$$\mathbf{A}_0=\begin{bmatrix}1-6\alpha_{x,0}\alpha_{z,0}\beta_{x,0}\rho_0/R_0 & 6\alpha_{x,0}\alpha_{z,0}\beta_{z,0}\rho_0/R_0\\ 3\beta_{x,0}(3\alpha_{z,0}^2-1)\rho_0/R_0 & -2+3\beta_{z,0}(3\alpha_{z,0}^2-1)\rho_0/R_0\end{bmatrix}. \tag{45}$$

The parameters entering the coefficient matrix $\mathbf{A}_0$ can be evaluated directly from the initial conditions of the motion. From standard formulas,[12] one has

$$\boldsymbol{\alpha}_0=\frac{\mathbf{v}_0}{v_0},\quad \boldsymbol{\beta}_0=\frac{\mathbf{v}_0\times(\mathbf{b}_0-g_0\mathbf{k})}{|\mathbf{v}_0\times(\mathbf{b}_0-g_0\mathbf{k})|}\times\frac{\mathbf{v}_0}{v_0}, \tag{46}$$

and

$$\rho_0=v_0^3/|\mathbf{v}_0\times(\mathbf{b}_0-g_0\mathbf{k})|, \tag{47}$$

where $\mathbf{b}_0$ is the initial value of $\mathbf{b}$ and it is assumed that $\mathbf{v}_0$ is non-vanishing. The corresponding quasilinear equations of motion are given by

$$\begin{Bmatrix}\ddot{x}\\ \ddot{z}\end{Bmatrix}+\omega_0^2\mathbf{A}_0\begin{Bmatrix}x\\ z\end{Bmatrix}=\begin{Bmatrix}b_x\\ b_z-g_0\end{Bmatrix}. \tag{48}$$

Solutions of the system (48) can be obtained directly by the use of the Laplace transform.

The quasilinear approximation represented by Eqs. (48) yields a transition from the linearized approximation (25) to the spherical approximation (14). The linearized equations assume that terms of the type $(x/R_0)^2$ in the Taylor expansion of $\mathbf{g}$ are negligible compared to terms of type x/R_0, which is equivalent to assuming $(s/R_0)^2\ll s/R_0$. From Eqs. (43) this condition requires that $\rho_0\ll R_0$, and one notes that Eqs. (48) reduce to the linearized equations (25) when $\rho_0\to 0$. The spherical solution (15) for motion in the x,z plane presupposes $\boldsymbol{\alpha}_0=\mathbf{i}$ and $\boldsymbol{\beta}_0=-\mathbf{k}$ at the origin; if note is taken of the constraining condition (12) on $\mathbf{b}_0$, Eq. (47) yields $\rho_0=R_0$. With these values of the parameters, Eqs. (48) reduce directly to the spherical equation, (14), of motion.

Since the terms in the Taylor expansion of $\mathbf{g}$ neglected by $\mathbf{g}_Q$ are of order s^3, and $s=O(t)$ when $\mathbf{v}_0\neq 0$, it follows that the absolute error $|\delta_Q\mathbf{r}|$ of the quasilinear solution is such that $|\delta_Q\mathbf{r}|=O(t^5)$ in general. Thus, the Taylor expansion of the quasilinear solution agrees with the true Taylor expansion of $\mathbf{r}$ through terms of order t^4 in general, as compared to the linearized solution (26), which agrees through terms of order t^3 at most. The quasilinear trajectory has contact of order not less than the fourth with the actual trajectory at the origin, and accordingly is an osculating approximation in general.

One can define error parameters for the quasilinear equations analogous to ξ and ζ in the linearized case, which fix surfaces of zero error for the component equations of (48). By this means one can show that the surfaces of zero error for both components coincide, and the quasilinear system (48) is exact if and only if the motion is constrained to the reference sphere S_0 (condition (12) holds), in which case the quasilinear system reduces to the spherical equation, (14). When $\rho_0\sim R_0$, the surfaces of zero error for both components are nearly in coincidence, and the quasilinear solution provides an approximation which is markedly superior to the linearized. When ρ_0 is small relative to R_0, the quasilinear trajectory differs only slightly from the linearized, and may yield only a small margin of gain. These results are illustrated by Fig. 8, which compares the quasilinear trajectories for the ballistic case ($\mathbf{b}=0$) with the corresponding ellipses and linearized trajectories when $\sigma_0=0$.

In passing, it can be noted that generalization of this quasilinear approximation to three dimensions is easily made (by assuming that the torsion of the curve of motion is likewise a slowly varying function).

V. LINEAR CONTINUATION

When significant variations of altitude above the earth occur on a trajectory of extended range, it is clear that none of the linear approximations given for $\mathbf{g}(\mathbf{r})$ are satisfactory in the large. In this case, the g-correction equation, (7), can always be solved by a step-by-step procedure in which the non-linear gravitational term over any interval of time is replaced by a constant value corresponding to the interval, as determined from the solution of the equation for times up to the interval in question. The solution in any interval is given by a suitable modification of the zero-order solution (11), and the solution is continued from interval to interval by identifying the initial conditions of one interval with the terminal conditions of the preceding interval. Such a method will be referred to as zero-order continuation, and, while always applicable, has the disadvantage of requiring a large number of steps for an extended range.

To reduce the number of steps required, the gravitational acceleration over any interval of time can be

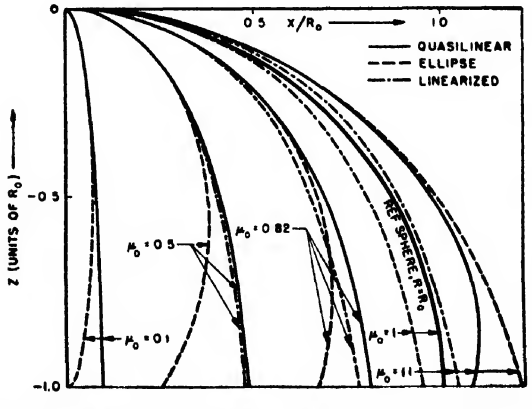


FIG. 8. Comparison of quasilinear trajectories with ellipses and linearized trajectories (ballistic case, $\mathbf{b}=0$).

TABLE I. Results of continuation solutions.

Solution	Step solution	Number of steps	Error $\|\delta r\|$ (mi.)
I	Linearized and quasilinear*	5	20
II	Linearized	39	46
III	Zero-order	39	254

* As specified in Fig. 3.

replaced by a linear approximation whose constants are determined by the solution up to the interval. The most advantageous approximation is the quasilinear. The initial conditions for any interval are again taken as the computed terminal conditions of the preceding interval. Such a method of solution will be referred to as linear continuation.

For simplicity, let the motion be plane with an initial point O_0 on the surface of the equivalent sphere S_e, and let the components of $\mathbf{b}$ be measured parallel to the axes x, z of the frame O_0. Let the total range, O to t_n of t, be divided into n intervals, t_{k-1} to t_k $(k=1, \cdots, n)$ with $t_0=0$. The computed terminal point O_{k-1} in the $(k-1)$th interval determines a frame $x_{(k)}$, $z_{(k)}$ with parameters R_{k-1}, Θ_{k-1} for the kth interval, as in Fig. 1 $(\Theta_0=0)$. In the frame O_{k-1} (which is introduced for computational convenience), the theory of the preceding sections for the frame O_0 can be applied to define the pertinent parameters g_{k-1}, ω_{k-1}, $\mathbf{A}_{k-1}$ by interchange of subscripts. If $\mathbf{B}_{k-1}$ designates the direction-cosine matrix

$$\mathbf{B}_{k-1}=\begin{bmatrix} \cos\Theta_{k-1} & \sin\Theta_{k-1} \\ -\sin\Theta_{k-1} & \cos\Theta_{k-1} \end{bmatrix}, \tag{49}$$

the g-correction equation in the kth interval when referred to the O_{k-1} frame is

$$\begin{Bmatrix} \ddot{x}_{(k)} \\ \ddot{z}_{(k)} \end{Bmatrix} + \omega_{k-1}^2 \mathbf{A}_{k-1} \begin{Bmatrix} x_{(k)} \\ z_{(k)} \end{Bmatrix} = \mathbf{B}_{k-1}' \begin{Bmatrix} b_x \\ b_z \end{Bmatrix} - \begin{Bmatrix} 0 \\ g_{k-1} \end{Bmatrix}, \tag{50}$$

where $\mathbf{B}_{k-1}'$ is the transpose of $\mathbf{B}_{k-1}$. The formulation (50) corresponds to the use of a quasilinear approximation, which reduces to a linearized or spherical approximation in the special cases

$$\mathbf{A}_{k-1}=\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \quad \mathbf{A}_{k-1}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{51}$$

respectively. The g-correction equation in the kth interval is then

$$\begin{Bmatrix} \ddot{x} \\ \ddot{z} \end{Bmatrix} + \omega_{k-1}^2 \mathbf{B}_{k-1}\mathbf{A}_{k-1}\mathbf{B}_{k-1}' \begin{Bmatrix} x \\ z \end{Bmatrix} = \begin{Bmatrix} b_x \\ b_z \end{Bmatrix} - \mathbf{B}_{k-1} \begin{Bmatrix} 0 \\ g_{k-1} \end{Bmatrix} \tag{52}$$

when referred to the O_0 frame.

In practice, it is found that linear continuation yields a considerable reduction in number of intervals as compared with zero-order continuation for the same terminal error. This fact appears by comparison of solutions I and II with solution III in Table I, which shows the results of three different continuation solutions carried out for the A10 trajectory. Solution I was based on the use of quasilinear approximations in the intervals (as specified in Fig. 3) where $\rho \sim R_0$; comparison of the results with those of solution II brings out the superiority of the quasilinear over the linearized approximation under this condition. In carrying out these computations, the $\mathbf{b}$-function (Fig. 4) was tabulated to four decimal places (after round-off) in units of g_e from the assumed trajectory (consisting of segments of analytic curves). Checks indicated that the rounding-off process in $\mathbf{b}$ introduced an error $|\delta\mathbf{r}|$ which amounted to <3 mi. for the solutions of Table I, which shows that the entries in the table for $|\delta\mathbf{r}|$ are significant for comparison of the approximations to $\mathbf{g}$.

It should be emphasized that the computations yielding Table I are of academic nature, due to neglect of the latitude-dependent component of $\mathbf{g}$ and of the centrifugal and Coriolis terms in the kinetic reaction of the g-correction equation. The sum of the first two of these acceleration terms varies over a range of $0.005g_e$ from the equator to the pole on the earth's surface. The systematic errors due to neglect of the variation of these terms on the A10 trajectory have been calculated approximately by a perturbation method for the case when the plane of motion coincides with the earth's equatorial plane (the error is then in this plane likewise). In this case the effect of variation of centrifugal reaction ($|\delta\mathbf{r}| \sim 30$ mi.) and of Coriolis reaction ($|\delta\mathbf{r}| \sim 700$ mi.) is extreme. These errors are larger by orders of magnitude than the upper limit ($|\delta\mathbf{r}| < 3$ mi.) to the round-off error found above for the data of Table I. Hence, the method of computation leading to Table I, in the particular form used, has only a limited practical validity for the A10 trajectory (except as an approximation in the roughly constant plane of motion for a trajectory in the neighborhood of a pole, where the Coriolis deflection is transverse).

It is clear from these considerations that the step solution in linear continuation must be modified on trajectories of extended range, in general, to include terms corresponding to the variation of the latitude-dependent component of $\mathbf{g}$, the variation of the centrifugal reaction, and the Coriolis reaction. The fact that these terms are considerably smaller than the main radial term of $\mathbf{g}$ materially simplifies the problem, which will not be attempted here because of the detail involved.

The advantage of linear continuation over numerical methods such as the Runge-Kutta process consists in the fact that the length of step or the integration interval is fixed by the variation of the slowly varying $\mathbf{g}$-function. From such error bounds as (22) or (31), one can estimate the number of steps required for a given terminal accuracy. In methods of the Runge-Kutta type, however, the integration interval is fixed by the

variation of the rapidly varying b-term, and the possible interval is considerably smaller. Linear continuation yields a reasonable compromise between the conflicting demands of a small number of steps and simplicity of the step solution.

In conclusion the authors wish to acknowledge many helpful discussions of this problem with Dr. W. C. Randels of Northrop Aircraft Corporation, Dr. R. Isaacs of The Rand Corporation, and Prof. T. Dantzig of the Rand consulting staff. Thanks are due also to Mrs. M. Irving and Mrs. J. Griffith for the computational work.

Diffraction Pattern in a Circular Aperture Measured in the Microwave Region

C. L. Andrews
General Electric Research Laboratory, The Knolls, Schenectady, New York
(Received February 2, 1950)

Measurements were made of the diffraction patterns of circular apertures from one to eight wave-lengths in diameter in the planes of the apertures and in the neighborhood of the apertures when a plane polarized electromagnetic wave was incident upon them. From Thomas Young's theory that the diffraction pattern is an interference pattern between the incident plane wave and wavelets from the edge of the aperture, the positions of maximum intensity have been predicted in the neighborhood of the aperture and the values of the intensities over the apertures checked with experiment.

INTRODUCTION

DIFFRACTION of light waves has been explained by Thomas Young as interference between the geometrically propagated incident wave and Huygen's wavelets caused by disturbance at the edge of the diffracting screen. Young's explanation was dropped promptly for Fresnel's theory that the diffraction pattern of a plane screen is the resultant of interfering wavelets from every point in any open portion of the infinitely extended plane of the screen. Lacking experimental facts, Fresnel made the simple assumption that the wave reaching the aperture of the screen was unperturbed. On the basis of Green's theorem and Fresnel's assumptions, Kirchoff developed a more exact theory of diffraction. A hundred years after Young and Fresnel, Rubinowicz[1] showed that Kirchoff's surface integral for the determination of the disturbance at a point on the opposite side of the aperture from the source may be transformed into two terms, one for the geometrically propagated wave through the aperture, and the other a line integral around the edge of the aperture. Thus Rubinowicz gave us two mathematically equivalent ways of thinking of diffraction by an aperture, one to think of intensity at a point in the diffraction pattern as the sum of all the effects from every point in the aperture and the other to think of diffraction as an edge effect.

Recently microwaves by their convenient length have provided a means of measuring the diffraction pattern in the plane of an aperture and on the sides toward and away from the source. It has been emphasized that the sharpest part of the diffraction pattern of an aperture is in the plane of the aperture itself.[2] Although this fact was recognized through measurements of microwave intensities in apertures, it could have been observed with light.

If the elementary apparatus for study of diffraction of light by a single slit is employed, the aperture may be opened to uncover a half dozen Fresnel strips relative to the point of observation. If the eye piece is then moved slowly toward the slit and at the same time the slit is diminished in width, so that the same number of dark bands is always observed, then, when the eye piece is finally focused on the plane of the aperture, sharp black lines are seen, each separated by a wavelength.

These simple observations with light are mentioned that the reader may confirm a fact of diffraction that has not been noted in the textbooks. The experimental work reported in this paper was performed with microwaves which provide the most precise means of measuring the diffraction patterns near apertures of the order of a few wave-lengths in diameter.

The assumption by Fresnel of constant intensity across the aperture, or more recent assumptions of constant intensity over most of a large aperture with tapering off only at the edge, have justification only in that they may be used in calculating correct intensities in the diffraction pattern beyond the aperture. The assumptions do not agree with experiment.

Thomas Young's theory of diffraction correctly predicts the positions of maximum and minimum intensity on the sides of the aperture toward and away from the source and in the plane of the aperture itself. Young assumed that the diffraction pattern of an aperture is the interference pattern between a geometrically propagated wave through the aperture and reradiated wavelets from the edge of the aperture. Thus the diffraction pattern of a slit is simply the interference pattern of three waves, the geometrically propagated wave through

[1] Rubinowicz, Ann. d. Phys. 53, 257 (1917).
[2] C. L. Andrews, Phys. Rev. 71, 777 (1947).

the slit and two cylindrical waves from the edges. The interference pattern of a plane wave and a cylindrical wave is a family of parabolas with their common principal focus at the edge, and their focal lengths an odd number of quarter wave-lengths. The interference pattern of the three waves in the plane perpendicular to the edges is not a set of curves but an array of points where the two families of parabolas intersect as shown in Fig. 1, the points where all three waves are in phase. Where the parabolas intersect at small angles the contours of intensity are elongated loops but near the aperture where the parabolas intersect nearly perpendicularly the contours are nearly circular and the gradients of intensity largest. Figure 1 gives the points of maximum intensity in the diffraction pattern of a slit three wave-lengths wide. Whether points A and B be positions of two long rods or the edges of a slit or the edges of a circular aperture upon which a plane polarized wave is incident, this simple theory of Thomas Young predicts closely the positions of maximum intensity in the magnetic plane between two parallel lines from the source through A and B. As the slit edges are separated the two families of parabolas move with them while the peaks of intensity move forward through the aperture from the standing wave region to the Fresnel region.

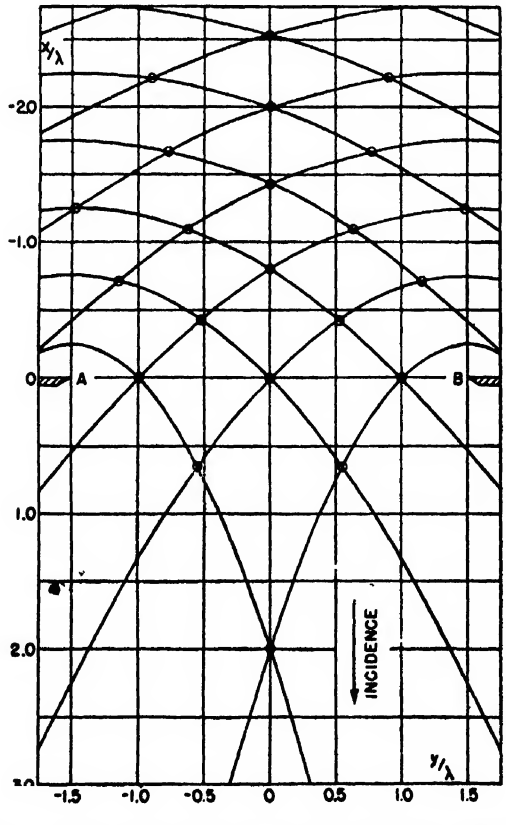


FIG. 1. Diagram for determination of positions of maximum intensity in the diffraction pattern in the magnetic plane of a circular aperture when a plane polarized wave is incident normally.

APPARATUS

The experimental arrangement was similar to that described previously.[1] The wave-length was reduced from 12.8 cm to 8.0 cm. A special triode oscillator was built for the purpose.[3] The parabolic reflector was 4 ft. (122 cm) in diameter. The equipment was set up out-of-doors on a roof to avoid reflections from walls and the distance from source to diffracting screen increased to 75 ft. (23 meters). The centers of the parabolic reflector and diffracting screen were 5 ft. above the roof. The block of the optical bench was driven by motor and screw at the rate of two wave-lengths per minute and the intensities recorded on a recording galvanometer (General Electric Photoelectric Recorder) with a sensitivity of 10 microamperes full scale and a period of 0.83 second.

There were two major sources of error (1) the lack of constancy of intensity in the cross section of the incident unperturbed beam, (2) the finite size of the antenna of the probe and the leads from the probe to the recorder. The automatic recorder made the search for sources of error simpler. The greatest variations of intensity of the unperturbed beam over a width of four wave-lengths were four percent and over a width of eight wave-lengths 10 percent. The crystal itself, Sylvania IN23B, with the brass tip of its capsule removed was used as dipole so that no coaxial lines would have to be employed to

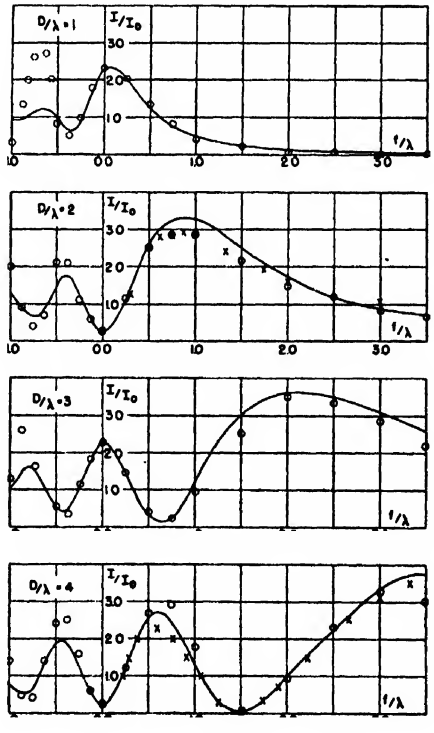


FIG. 2. Diffraction patterns along axes of circular apertures 1, 2, 3, and 4 wave-lengths in diameter as computed from Kirchoff's theory. Circles are experimental data. Crosses are data of Severin.

[3] C. L. Andrews, General Electric Review 50, 40 (1947).

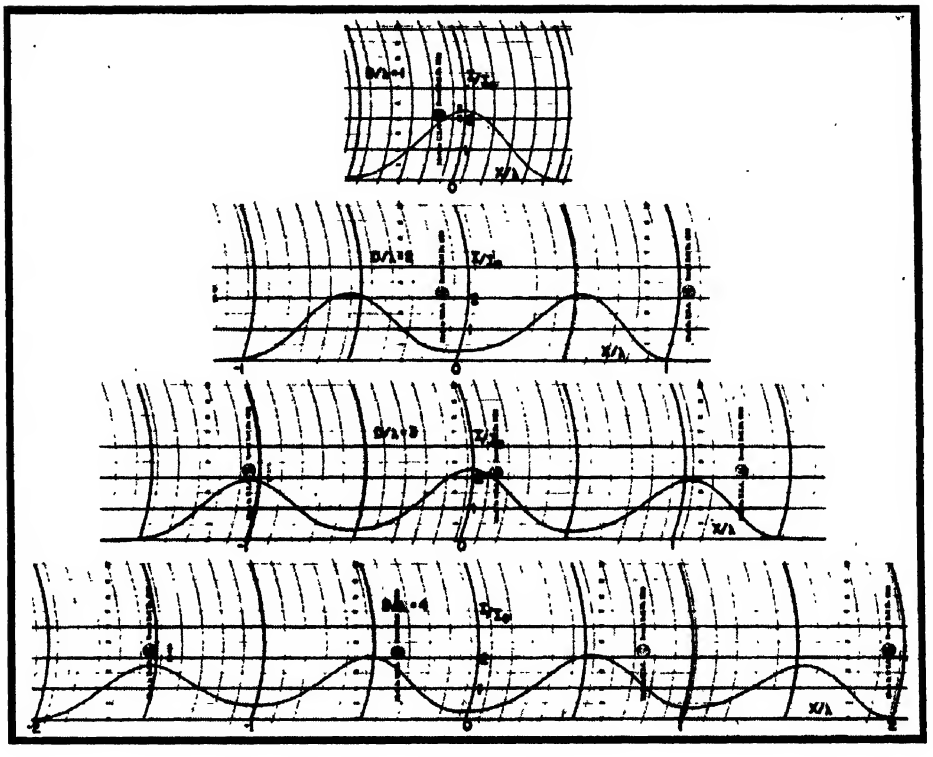


FIG. 3. Recorded graphs of intensity of radiation along magnetic diameters of circular apertures 1, 2, 3, and 4 wave-lengths in diameter.

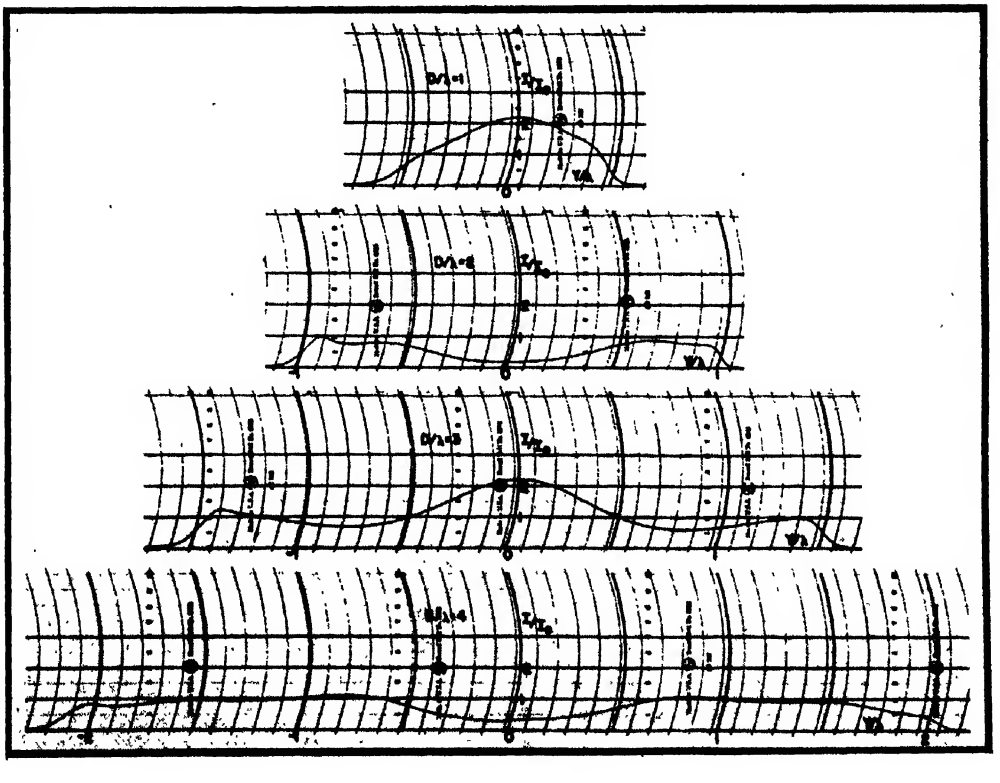


FIG. 4. Recorded graphs of intensity of radiation along electric diameters of circular apertures 1, 2, 3, and 4 wave-lengths in diameter.

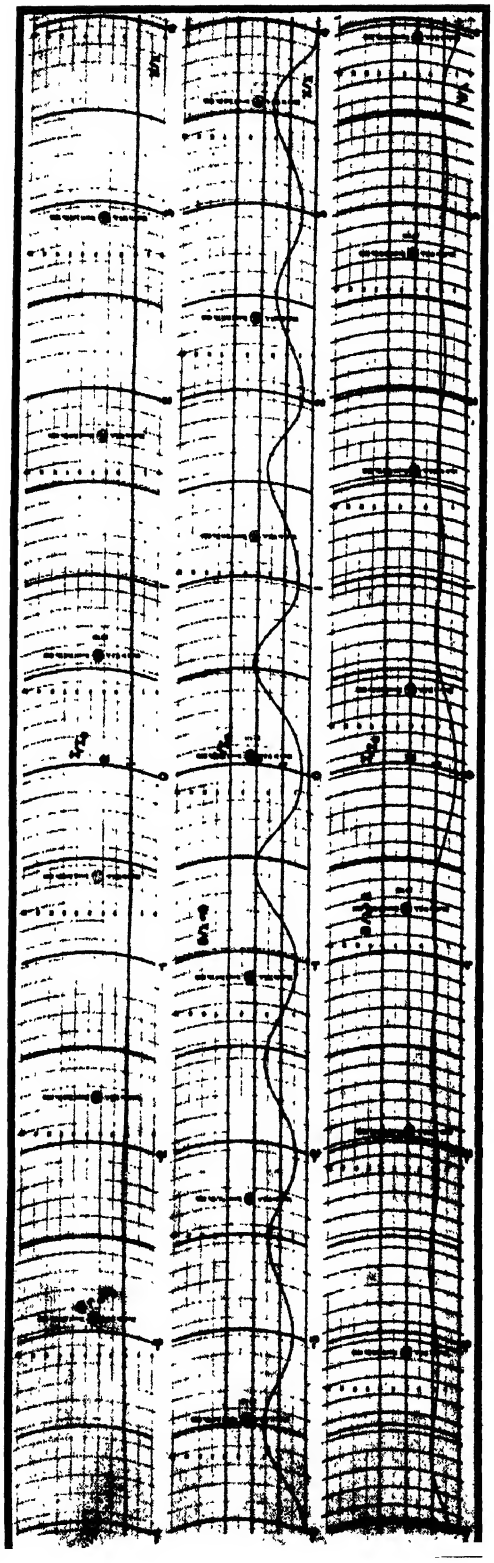


FIG. 5. Intensity of radiation of the unperturbed beam and along magnetic and electric diameters of circular apertures eight wave-lengths in diameter.

disturb the field. A fine twisted lead of No. 23 wire was attached with soft solder to the ends of the capsule of the sealed silicon crystal and extended in the direction of propagation for three meters and guyed by strings and rubber bands. Thence it extended out of the beam perpendicularly to the electric field.

The errors caused by the finite size of the antenna were estimated in two ways:

(1) By comparing records of the same diffraction patterns as the length of the probe was reduced.

(2) By studying the diffraction patterns of rods varying from 0.5 to 0.1 wave-length.

When the antenna was reduced from 0.25 to 0.20 wave-length, the patterns were not altered noticeably except when the probe was closer than a quarter wave-length to the edge of an aperture. Likewise, it was noted that rods of 0.2 wave-length long had negligible effects on the field at distances greater than one-quarter wave-length.

It was estimated that the greatest source of error in all patterns shown in the following graphs was the variation of intensity in the cross section of the unperturbed beam except at points within an eighth wave-length of an edge on the H diameter or a quarter wave-length of an edge on the E diameter where the largest error was caused by the finite size of the probe.

EXPERIMENTAL RESULTS

In Fig. 2 the circles indicate experimental data of intensity of radiation I relative to the intensity I_0 of the unperturbed beams taken along the axes of circular apertures in metal screens 1, 2, 3, and 4 wave-lengths in diameter. f/λ is the distance in wave-lengths along the axis measured from the plane of the aperture in the direction of propagation. The crosses are experimental data taken from the results of Severin[5] for wave-lengths of 6 and 10 cm. Severin coated the front and back of his metal screen with a "paint" that reduced reflections to 10 percent of the metallic reflection. Near the aperture data are the same for the reflecting and absorbing screen. The curves of Fig. 2 are plots of the calculations from Kirchoff's theory.[6] Note that the best agreement with experiment is at the center of the aperture. If the assumptions of Kirchoff's theory were not conflicting and if the solution were correct then the solution should yield the original assumption of an unperturbed beam in the plane of the aperture. Therefore, we shall ignore the origins of Kirchoff's expression and consider it as a semi-empirical formula. The disagreement with experimental data on the side toward the source may be expected because of reflection from the mirror surface on the back side of the screen.

Note that when the aperture is an odd number of wave-lengths in diameter I/I_0 at the center is 2.25. When the diameter is an even number of wave-lengths, I/I_0 is 0.25.

[5] H. Severin, Zeits. f. Naturforshung, I 9, 487–495 (1946).
[6] See reference 1, Eq. (4) and Fig. 1.

Figure 3 is of recorded graphs of the intensity along the magnetic diameters of circular apertures 1, 2, 3, and 4 wave-lengths in diameter when a plane polarized beam is incident normally on the screen.

Figure 4 is a similar record of intensities along the electric diameters. Note that the measurements along the electric diameters are not dependable closer than a quarter wave-length from the edge. It may be noted that whatever the length of the dipole probe the recorded intensity begins to drop as soon as one end of the dipole probe goes behind the screen and drops to zero when the probe is completely hidden from the source.

Figure 5 is a recording of the intensity of the unperturbed beam and the intensities along the magnetic and electric diameters of an aperture eight wave-lengths in diameter.

A search for a meaningful empirical expression to describe the patterns resulted in

$$(u/u_0)_v = 1 - \frac{1}{2\pi} \oint \epsilon^{i\beta} \cos\theta \cos\gamma (ds/r), \qquad (1)$$

$$(u/u_0)_h = -\frac{1}{2\pi} \oint \epsilon^{i\beta} \cos\theta \sin\gamma (ds/r), \qquad (2)$$

u is the complex amplitude of the resulting wave at point P of Fig. 6 in terms of u_0, the amplitude of the incident wave. The subscripts v and h indicate vertical and horizontal components. The term unity is for the vertically polarized incident beam. The negative sign indicates the reversal of phase of the reradiated wavelet from the edge. ds is an element of edge and r its distance from point P. β is the phase angle, $2\pi r/\lambda$. Multiplication by $\cos\theta$ resolves the component of the amplitude of the incident wave parallel to the element of edge ds. $\cos\gamma$ and $\sin\gamma$ resolve the amplitude of the wavelet from ds arriving at P into vertical and horizontal components. Since the expression contains ideas clearly expressed by Thomas Young, we shall call it Young's Circuital Form for determining the diffraction pattern in the plane of an aperture in a flat screen when a plane polarized transverse wave is incident normally on the plane of the aperture.

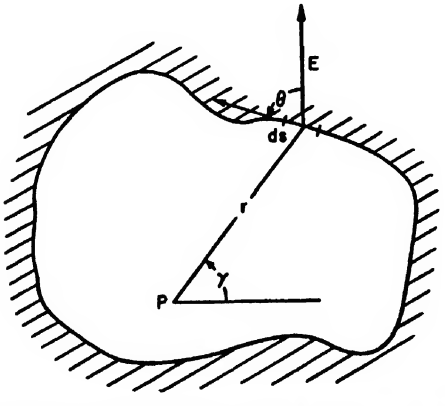


FIG. 6. Aperture in screen upon which a vertically polarized beam is incident normally.

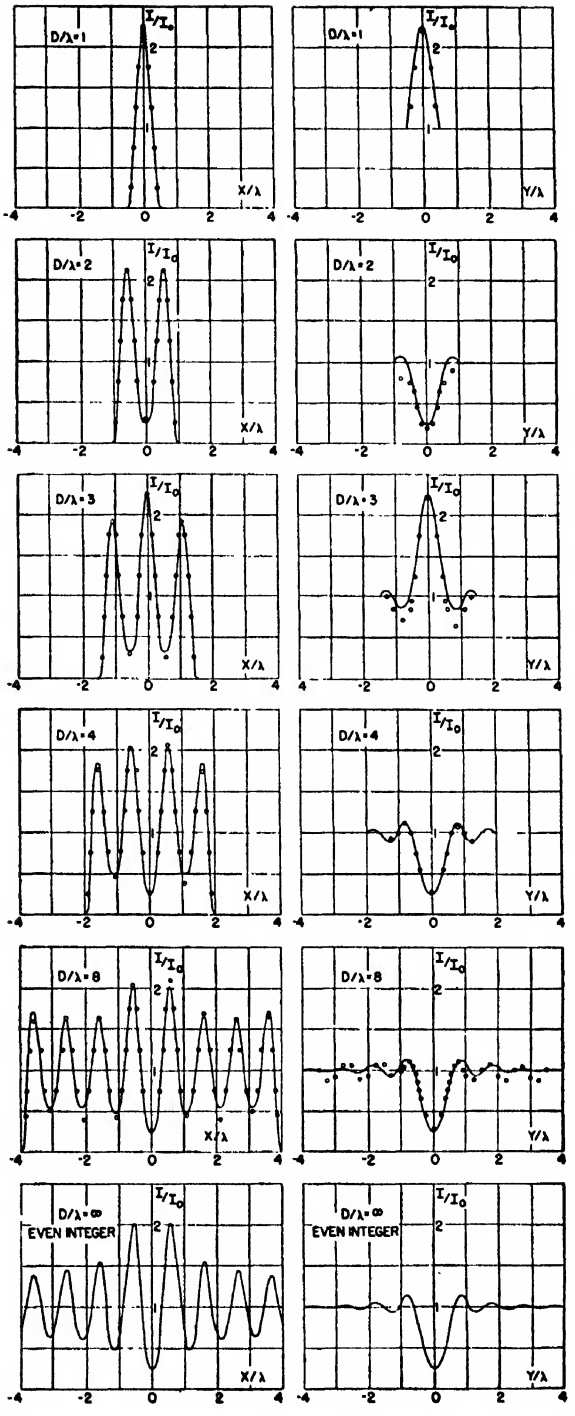


FIG. 7. Computed intensities of radiation along magnetic and electric diameters of circular apertures 1, 2, 3, 4, 8 and an infinite even integer number of wave-lengths in diameter. The circles indicate experimental data from Figs. 3, 4, and 5.

If the aperture is a circle of radius R and x and y are

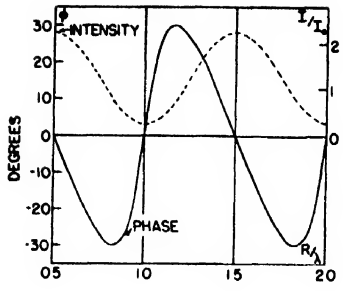


FIG. 8. Phase and intensity of radiation at center of circular aperture *versus* diameter.

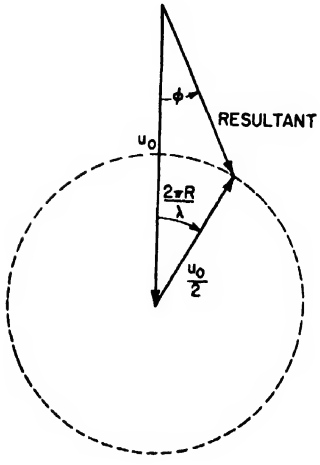


FIG. 9. Vector diagram for determination of phase and amplitude at center of circular aperture one wave-length in diameter or greater.

the cartesian coordinates taken along the horizontal and vertical diameters from the center to express the position P, then

$$(u/u_0)_v = 1 - \frac{1}{2\pi} \oint \epsilon^{i\beta} (R/r)^2 (\cos^2\theta - (x/R)\cos\theta) d\theta, \quad (3)$$

$$(u/u_0)_h = -\frac{1}{2\pi} \oint \epsilon^{i\beta} (R/r)^2 \times (\sin\theta\cos\theta - (y/R)\cos\theta) d\theta. \quad (4)$$

All of the independent variables may be expressed in terms of any one of them. We were not able to integrate the expressions but evaluated the integrals for points along the magnetic and electric axes by dividing the circle into ten degree steps and adding vectorially. Some cases were checked by dividing the aperture into two degree steps.

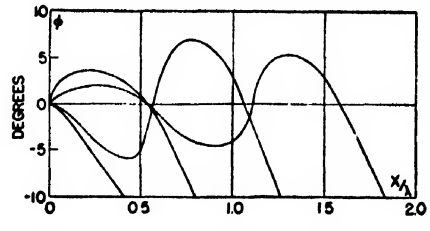
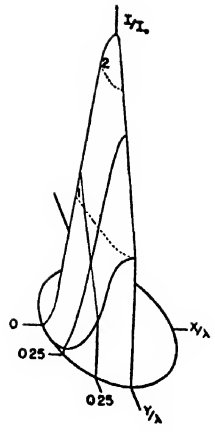


FIG. 10. Phases along the magnetic diameter of apertures 1, 2, 3, and 4 wave-lengths in diameter.



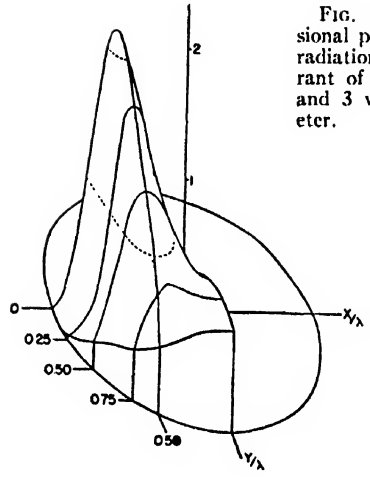
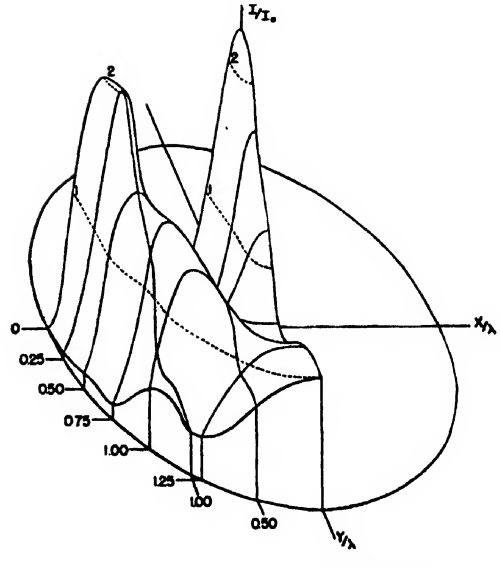


FIG. 11. Three dimensional plots of intensity of radiation over one quadrant of apertures of 1, 2, and 3 wave-lengths diameter.



The two columns of curves in Fig. 7 are of the computed intensities of radiation along magnetic and electric diameters of circular apertures 1, 2, 3, 4, 8 and an infinite even integer number of wave-lengths in diameter. The circles indicate experimental data from the curves of Figs. 3, 4, and 5.

Two features other than the general agreement of the data and calculations are to be noted. For both the experimental and calculated curves (1) the peaks are narrower than the valleys (2) the maxima and minima do not lie at exactly integer half wave-lengths from the center but slightly beyond those positions.

At the center of the circular aperture the phase ϕ alternates with increasing radius between 30° and −30° by a saw-toothed relation indicated by Fig. 8. A plot of I/I_0 *versus* R/λ is given with coordinates on the right to indicate that, when the intensity differs most from that of the incident beam, the phase is the same as that of the incident beam.

At the center

$$I/I_0 = 5/4 - \cos 2\pi R/\lambda, \tag{5}$$

$$\tan\phi = \frac{\frac{1}{2}\sin 2\pi R/\lambda}{1-\frac{1}{2}\cos 2\pi R/\lambda}. \tag{6}$$

Equations (5) and (6) are seen to follow from the vector diagram of Fig. 9.

Equations (5) and (6) are a special case of Kirchoff's expression[6] for the amplitude and phase on the axis provided $R/\lambda > \frac{1}{2}$. For smaller radii the point of observation may be chosen on the axis near the center on the side of the screen away from the source. Small apertures have not been treated experimentally. They will require a smaller probe than the cases reported here.

Figure 10 is a plot of phases along the *II* diameter for diameters of 1, 2, 3, and 4 wave-lengths. The calculations of these angles may be in error by as much as 0.5 degree. The phase angles are small and where the intensity is a maximum the phase is the same as that of the incident beam. The largest phase angles occur where the intensity is approaching zero at the edge. Thus with good approximation one may assume constant phase over the aperture when applying Huygens' principle to calculate the diffraction pattern beyond the aperture. Smythe's[7] method of the double current sheet will be readily applicable. However, the approximation of constant phase is justified only when the diameter of the aperture is an integer number of wave-lengths.

If, for apertures an integer number of wave-lengths wide, one considers the positions of maximum intensity, Fig. 8, as line sources, then the Fraunhaufer diffraction pattern is easily calculated vectorially and is the same as that calculated by zone or strip methods from the assumption of constant phase and intensity over the aperture.

Figure 11 is of three-dimensional plots of intensity over one quadrant of apertures of one, two and three wave-lengths in diameter from experimental data.

CONCLUSION

In view of these experimental observations the classical treatment of diffraction is in need of review. Although there have been many volumes of theory written on diffraction, there have been comparatively few experimental results reported in sixty years. Microwaves now provide a precise means of measurement of diffraction patterns near apertures.

Young's theory of diffraction yields the diffraction pattern in the plane of a circular aperture.

ACKNOWLEDGMENT

It is a pleasure to acknowledge the support and helpful discussions throughout this study of W. C. White and E. D. McArthur of the General Electric Research Laboratory.

[7] W. R. Smythe, Phys. Rev. **72**, 1066 (1947).

The Retrograde Motion of the Arc Cathode Spot

C. J. GALLAGHER
General Electric Research Laboratory, Schenectady, New York
(Received February 6, 1950)

The motion of the cathode spot in a direction opposite to that predicted by Ampère's law depends on the arc current, gas pressure, kind of gas, and magnetic field strength. Studies of the effects of these variables have been made, including measurements of velocity and the critical pressure at which reversal of motion occurs. The phenomena observed have not been clearly explained by any of the pictures presented to date. The existence of the retrograde motion indicates very strongly that the positive-ion space charge outside the cathode is all important in determining the mechanism of current transfer.

IN the low pressure arc the cathode spot, when subjected to a transverse magnetic field, may move in a direction opposite to that which Ampère's law predicts.[1,2] The direction and speed of this retrograde motion are determined by: (1) gas pressure, (2) magnetic field strength, (3) arc current, and (4) kind of gas. It is not certain at this writing whether or not the cathode material has any appreciable effect. This paper will describe some experiments in which the above variables were investigated as to their effect on the motion.

The first series of experiments is concerned with a low pressure mercury arc, using a mercury pool cathode. The geometry of the tubes used was such that the cathode was ring-shaped. The tube was placed in the gap of an electromagnet with an annular structure as shown in Fig. 1. The magnetic field was radial, so that the current entering the cathode spot was always perpendicular to the field. Both tube and magnet were immersed in oil, which was cooled by circulating through a heat exchanger to provide a steady stream of oil cooler than the bath itself. This stream was directed against the wall of the tube and thus determined the condensing temperature of the mercury inside. The condensing temperature could be considered the same as that of the stream, except for the small thermal drop in the glass wall of the tube. The tube was operated at a line voltage of 250 v d.c. with a variable series resistance to regulate the current.

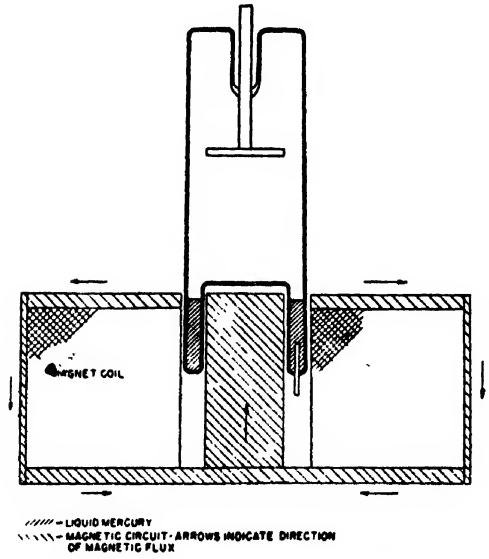


FIG. 1. Mercury pool tube in radial magnetic field.

The velocity of the spot was obtained from the frequency of rotation. The light from the region of the spot was transmitted through the oil by a glass rod and then allowed to fall on a photo-multiplier tube, the output of which was applied to the vertical plates of an oscilloscope. A sine wave of known frequency was applied to the horizontal plates and varied until a one to one Lissajou pattern was obtained. The method was not effective below about 20 c.p.s. because of random fluctuations in the position of the path of the spot over the mercury surface. At high speeds, the spot stayed against the wall of the tube so that very regular patterns could be obtained.

The velocity of the spot as a function of magnetic field, arc current, and pressure is shown in Figs. 2, 3, and 4, respectively. As shown in Fig. 2, the velocity varies almost linearly with magnetic field over the middle range, with an exponential-like variation at low fields. The slopes of the linear portions of the curves appear to be independent of current. At high fields, above 3000 gauss, the results are somewhat uncertain. It was found that the points could not be reproduced with any great accuracy. The voltage across the arc at high magnetic fields is high and may cause instability, although the arc voltage did not show any unusual fluctuations in this region such as might be expected if the arc were unstable. In Fig. 3, the same data are replotted, but with arc current as the independent variable. There is a definite leveling-off point for each value of field, with the constant velocity region setting in at smaller values of current as the field is increased.

The velocity of the spot depends on the pressure, decreasing as the pressure increases. In fact, if the pressure is sufficiently increased, the spot can be made to slow down to zero velocity and then reverse its motion. The pressure at which reversal occurs depends on the magnetic field and the current, as shown in Fig. 4. The reversing pressure as a function of magnetic field is shown for different arc currents. In general, less field

[1] N. Minorsky, J. de phys. et rad. 9, 127 (1928).
[2] C. G. Smith, Phys. Rev. 62, 48 (1942).

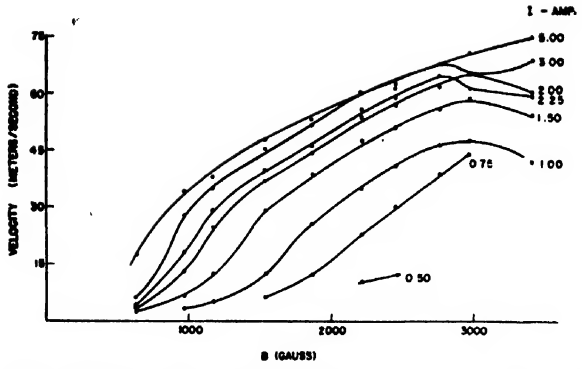


FIG. 2. Velocity of cathode spot on Hg pool *vs.* magnetic field for various arc currents. Pressure = 0.7μ.

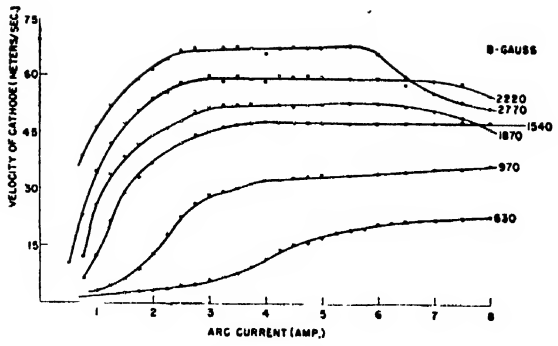


FIG. 3. Velocity of cathode spot on Hg pool *vs.* arc current for various magnetic fields. Pressure = 0.7μ.

is required to reverse the spot at high currents than at low. This suggests that the point of reversal represents a condition of balance between the forces acting on the spot and the forces acting on the column, since it has been observed that the column always tries to move in the "correct" direction.

The appearance of the rotating discharge is shown in the series of pictures in Fig. 5. These are moving pictures, taken at high speed (4000 frames per second). The arc was moving at a speed of about 75 meters per second, with a current of 3.5 amp. The column can be seen extending from the spot away from the direction of motion. The lower image is due to reflection in the magnet surface. The pictures reveal an interesting phenomenon in that occasionally there appears a large dark space between the column and the glow at the cathode. Evidence of this is beginning to appear in the second frame, with a larger gap showing in the third. The dark region has been filled with glowing gas again in the two final pictures. Examination of the complete film indicates that these gaps are randomly distributed in time with no apparent periodicity. Note that the glow of the cathode persists, regardless of the column behavior.

The phenomenon of retrograde motion and reversal was observed with many other gas and electrode combinations. Gases used included nitrogen, oxygen, hydrogen, carbon dioxide, helium, and argon, with electrodes of copper, carbon, tungsten, molybdenum, aluminum, nickel, and cadmium. There was found to be very little, if any, effect of electrode material on the reversing pressure. The principal effect of the electrode appeared to be in determining the stability of the arc. It was not possible to establish arcs at these currents between electrodes which had been carefully cleaned and outgassed. It was necessary in all cases to operate with metals which were oxidized to some extent.

The reversing pressure was sharply affected by the kind of gas used. The results obtained are shown in Fig. 6. The arc current was adjusted to 4 amp. and the magnetic field kept at a low value of about 100 gauss. There is a wide range in reversing pressures, ranging from about 1 cm for mercury to 16 cm for helium, with other gases lying in between. As Fig. 6 shows, there is some correlation between the reversing pressure and the first resonance potential of the gas. The first resonance potential may be taken as a rough measure of the voltage drop between the cathode and the column (the "cathode drop"). The first resonance potential is probably a better choice than the ionization potential because of the possibility of ionization occurring in several steps rather than a single ionizing collision. The probability of step-wise ionization is high where the current density is large. The current density at the arc cathode spot may be greater than 50,000 amp. per cm².[3]

It will be noted that mercury exhibits the lowest reversing pressures. For every other gas investigated, the reversing pressures were higher. In fact, retrograde motion can be observed at pressures as high as atmospheric, depending on the kind of gas, the current, and magnetic field strength. In argon with oxidized tungsten electrode, as shown in Fig. 7, the reversing pressure can be as high as 60 cm Hg, and as low as 5 cm Hg. It will be noted that the effect of current and magnetic field on the reversing pressure is, in general, the same as found in mercury, i.e., the reversing pressure is greater fields or smaller currents.

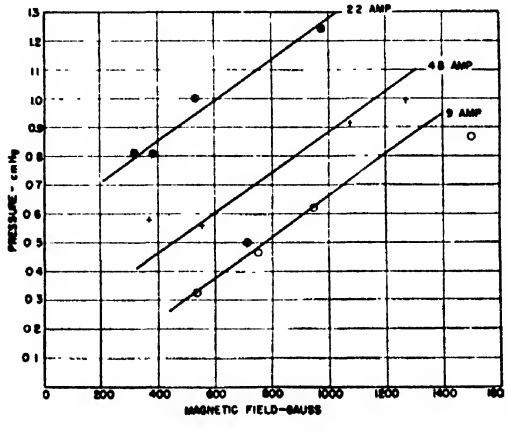


FIG. 4. Reversing pressure *vs.* magnetic field in mercury pool tube.

[3] J. D. Cobine and C. J. Gallagher, Phys. Rev. 74, 1524 (1948).

There is at present no adequate explanation for the phenomena described here. A complete theory would undoubtedly require a knowledge of the mechanism of current transfer at the arc cathode, since the action of the magnetic field must be either at the cathode surface or else at the region of high ion density immediately outside it. If one rules out the Righi-LeDuc effect, as originally proposed by Smith,[2] and other similar effects (Hall, Ettingshausen, and Nernst effects,[4] there remains only the effect of the magnetic field on electron trajectories as the primary physical phenomenon. The electrons involved are probably those emitted by the cathode. The direct effect of the magnetic field on the paths of positive ions is probably quite small, at least in comparison with the change in electron paths. However, the electron trajectories are bent in a direction opposite to the retrograde motion, so that the motion itself is a secondary effect. A simple calculation of the motion of a single electron in crossed electric and magnetic fields indicates that the particle is deflected only a small amount before colliding. This is especially true at higher pressures where the mean free path is short. Bending the electron paths would certainly shift the center of gravity of the positive ion space charge outside

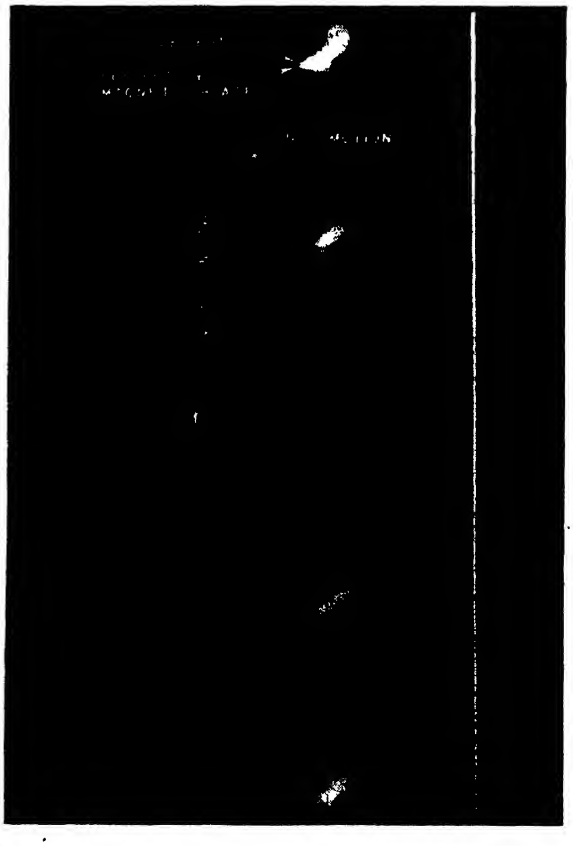


FIG. 5. High speed pictures of cathode spot moving on Hg pool (4000 frames per second).

[4] P. W. Bridgman, *The Thermodynamics of Electrical Phenomena in Metals* (The Macmillan Company, New York, 1934).

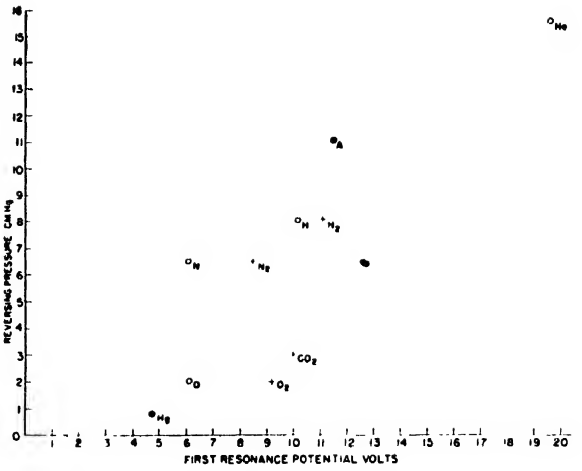


FIG. 6. Reversing pressure *vs.* resonance potential. arc current—4 amp., magnetic field—100 gauss.

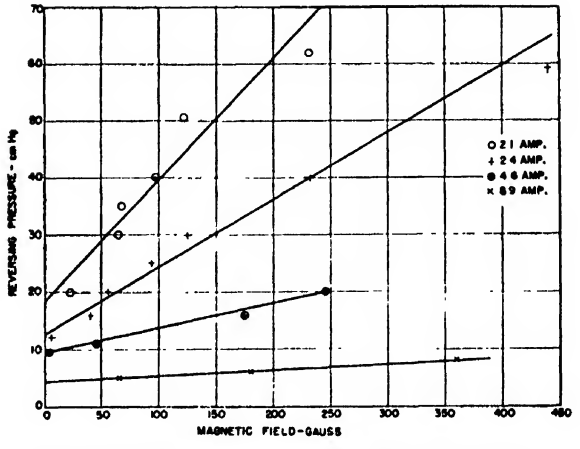


FIG. 7. Reversing pressure *vs.* magnetic field oxidized tungsten in argon.

the cathode, since these ions are probably formed by collisions with emitted electrons. Longini[5] makes a plausible case for a net space-charge shift in the retrograde direction by considering that the positive-ion space charge is shifted less than the electron space charge. It is not clear from his picture how reversal can occur, or why the direction of motion depends on the pressure, the current, or the magnetic field. Himler and Cohn[6] have offered a less plausible explanation by considering the electrons to be emitted from the cathode in all directions, with those emitted in the "correct" direction being returned to the cathode without contributing to the current. Those emitted in the retrograde direction would produce a positive-ion space charge in that direction which would produce a field sufficient to extract more electrons from the cathode. It does not seem possible that there can be any such

[5] R. L. Longini, Phys. Rev. 72, 184 (1947); 71, 642 (1947).
[6] G. J. Himler and G. I. Cohn, Elec. Eng. 67, 1148 (December, 1948).

random distribution in directions of emission, at least, not with the spread that this picture would require.

Both of the above pictures point to the position of the positive-ion space charge as the important factor in determining the position of the cathode spot which in turn means that the emission process depends on the positive ions. That this must be so seems to be borne out by two experimental observations. First, there is the fact that the velocity of the spot is of the order of the probable thermal velocity of the gas atoms (or ions). In fact, it was not possible to get velocities quite as high as thermal velocity. Unless the electric fields are very distorted, the motions of the ions parallel to the cathode will be determined by the thermal velocities. A second evidence of the importance of the position of the positive ions is the observations by Smith,[7] and also by Cobine and Gallagher,[8] that the cathode spot does not exhibit retrograde motion if the cathode is heated to a temperature high enough to support thermionic emission. In this case, the positive-ion space charge is no longer required except to accelerate electrons which have already been emitted, so that the position of the space charge is less important in determining the positions of the cathode spot. These observations are not conclusive, but they do indicate that the positive-ion space charge plays an important role in determining the motion of the spot, and, hence, in maintaining current transfer at the cathode surface.

[7] C. G. Smith, Phys. Rev. 73, 543 (1948).
[8] J. D. Cobine and C. J. Gallagher, Elec. Eng. 68, 469 (1949).

Transonic Potential Flow of a Compressible Fluid*

W. R. Sears
Cornell University, Ithaca, New York
(Received February 20, 1950)

Even under the assumptions of irrotational, isentropic flow, which have been found generally useful for subsonic and supersonic cases, the equations of gas flow are relatively intractable for mixed, transonic situations. Approximate methods of solution used by a number of investigators are reviewed briefly, as well as the hodograph technique, which yields exact solutions of the equations for plane flow. It is pointed out that all the methods predict smooth, potential mixed flows involving imbedded regions of supersonic speed and both acceleration and deceleration through the speed of sound. There is no experimental verification of the existence of such flows.

Three possible explanations for this sharp discrepancy between experiment and perfect-fluid theory have been advanced; namely, (a) effects of viscosity, (b) non-existence of neighboring solutions, and (c) temporal instability. These are reviewed in turn. None has led to a complete explanation, to date. Kuo's stability calculations are described briefly. His results indicate that stable, smooth, mixed flows may exist if certain conditions are satisfied.

1. METHODS OF CALCULATING TRANSONIC POTENTIAL FLOW

MUCH of our progress in the theory of compressible fluid flows is based on the compatible assumptions of irrotational flow and isentropic changes of state. These approximations are justifiable, in many situations, because a gas of small viscosity in purely subsonic flow, or in flow involving only straight or weak shocks, does flow nearly isentropically and irrotationally. Thus, for many subsonic and supersonic flows, the results of potential-flow theories are in good agreement with experimental observations. We are concerned here with the validity of those assumptions in cases of transonic flow.

Even for steady plane flows that are inviscid, isentropic, and irrotational the differential equations that describe the flow patterns are non-linear, and therefore do not yield, directly, to standard and general mathematical methods of solution. To be sure, these equations can be made linear, by the simple approximation of neglecting the non-linear terms, and are still useful to describe situations occurring at low speeds and again at moderately high speeds, but not near the speed of sound. When the gas takes on speeds close to the sonic speed, relative to a solid obstacle or the walls of a channel, the same essential non-linearity appears in the equation of continuity that is so clearly demonstrated at the throat of a Laval nozzle. The familiar "linearized theory" then fails.

The engineer has not been completely unable to cope with this difficulty. Recognizing the essential non-linearity and the formidable mathematical difficulty of his problem, he has, with some success, taken recourse to such dependable methods as purely numerical computation and series expansion.

G. I. Taylor,[1] for example, in 1930 assumed the velocity potential to be expanded in a double power series in the Cartesian coordinates x and y,

$$\phi = a_1x + a_2x^2 + c_2y^2 + a_3x^3 + c_3xy^2 + a_4x^4 + c_4x^2y^2 + e_4y^4 + \cdots. \quad (1)$$

* An invited paper delivered at the meeting of the American Physical Society at Columbia University, February 4, 1950.

[1] G. I. Taylor, "The flow of air at high speeds past curved surfaces," Aero. Res. Comm. Reports and Memo. No. 1381, London (1930).

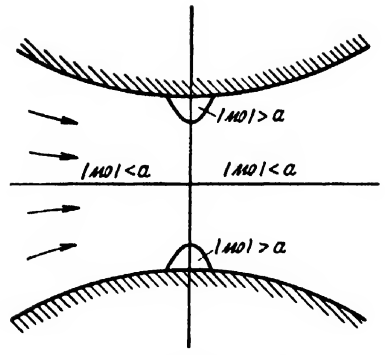


FIG. 1. Transonic potential flow through nozzle, according to calculation of Taylor (see reference 1).

He then proceeded to calculate numerical solutions to the non-linear equations, for the case of plane flow through a symmetrical nozzle. This procedure led to solutions involving both subsonic and supersonic speeds, which will be described later.

Using an entirely different approach, following Prandtl and Ackeret, a number of writers have imagined the velocity potential (or the stream function) to be expanded in powers of a thickness parameter τ, for a family of flow patterns produced around a family of slender cylindrical objects:

$$\phi = U\{x + \tau\phi_1 + \tau^2\phi_2 + \cdots\}. \qquad (2)$$

For example, the parameter τ might be the thickness ratio, for a family of symmetrical airfoils, or the camber, for a family of circular-arc profiles. This results in an iteration procedure by which the successive coefficients of τ^n are calculated from the knowledge of the preceding ones, and in fact the first approximation is precisely the result of the "linearized," or Prandtl-Glauert theory mentioned above. Although the convergence of this process has never been established theoretically, the numerical work appears to converge very satisfactorily, at least at low Mach numbers, and there is no apparent difficulty in producing flows of the transonic type; that is, having both supersonic and subsonic local speeds. The most important work along this line was done by Carl Kaplan[2] in America and by H. Görtler[3] in Germany.

Howard Emmons,[4] in 1946, undertook to attack the problem purely numerically. Approximating to the differential equations of plane motion by differences, he applied the relaxation technique successfully, making the switch from subsonic to supersonic conditions when necessary, and finally, if necessary, introducing shock waves.

[2] Carl Kaplan, "The flow of a compressible fluid past a curved surface," N.A.C.A. Tech. Report No. 768 (1943).

[3] H. Görtler, "Gasströmungen mit Übergang von Unterschall- zu Überschallgeschwindigkeiten," Zeits. f. Angew. Mat. Mech. **20**, 254–262 (1940).

[4] H. W. Emmons, "The theoretical flow of a frictionless, adiabatic, perfect gas inside of a two-dimensional hyperbolic nozzle," N.A.C.A. T.N. 1003 (1946). Also, "Flow of a compressible fluid past a symmetrical airfoil in a wind tunnel and in free air," N.A.C.A. T.N. 1746 (1948).

More recently, Theodore von Kármán,[5] and, independently, G. Guderley,[6] discovered that a perturbation theory of a different sort, again for slender bodies, could be made for the transonic regime. In this theory it is assumed that every velocity represents a small perturbation of a uniform parallel stream at sonic speed a^*. The resulting equation for the perturbation velocity

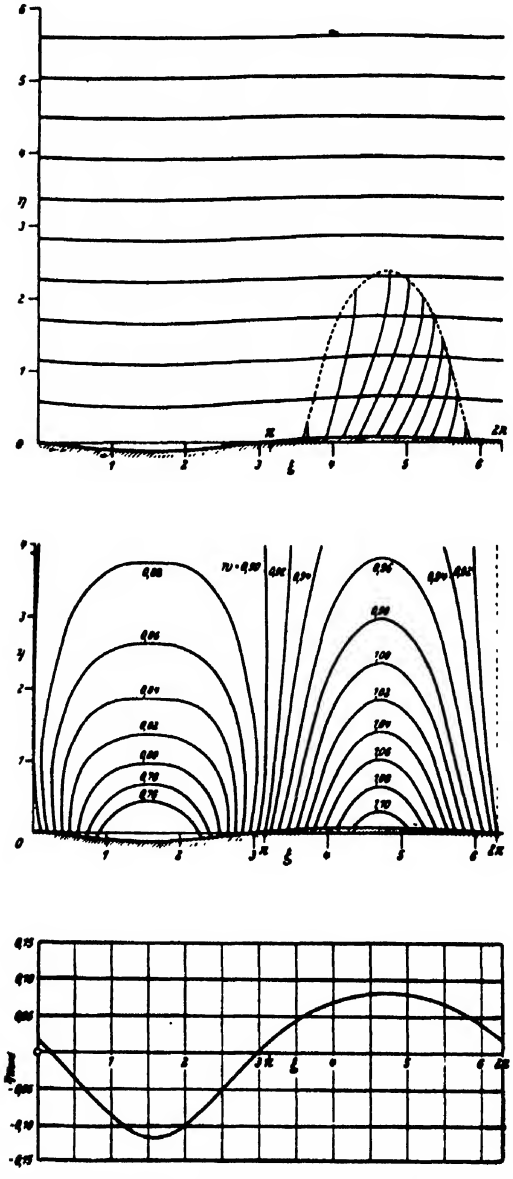


FIG. 2. Transonic potential flow over a wavy wall, according to calculations of Görtler (see reference 3). Top: streamlines and (in supersonic region) Mach lines. Center: curves of constant Mach number. Bottom: shape of the wall, ordinate enlarged ten times.

[5] Th. von Kármán, "The similarity law of transonic flow," J. Math. and Phys. **26**, 182–190 (1947).

[6] G. Guderley, "On the transition from a transonic potential flow to a flow with shocks," U.S.A.F. A.M.C. Tech. Report No. F-TR-2160-ND (1947).

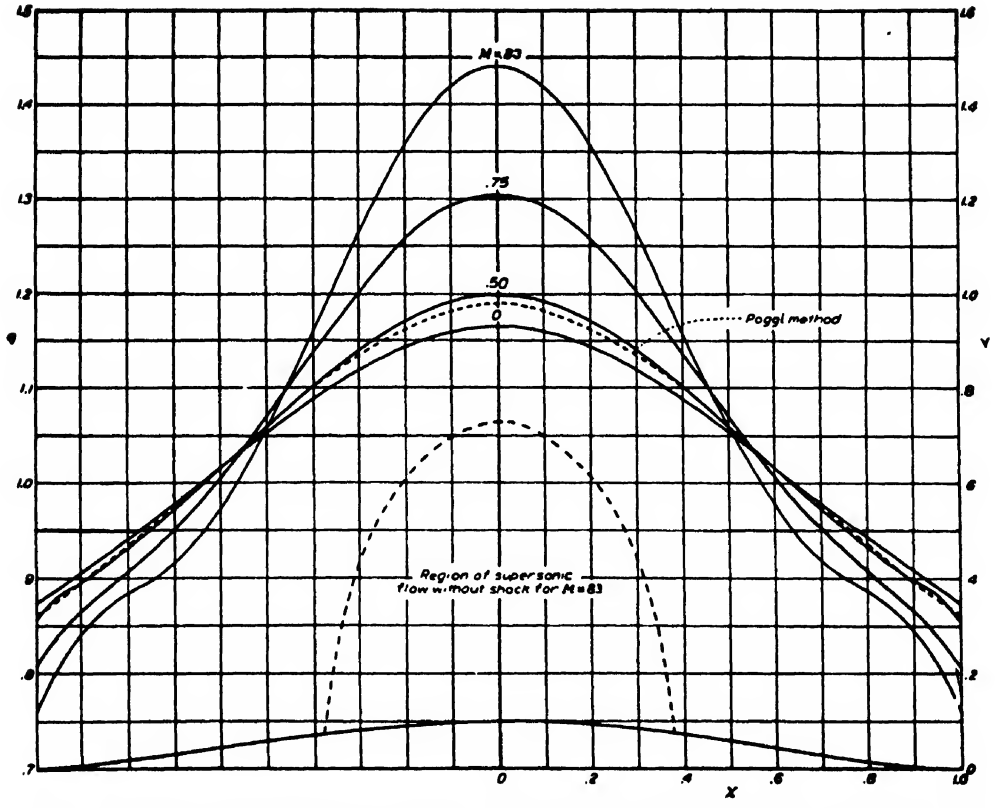


FIG. 3. Transonic potential flow over a bump, according to calculations of Kaplan (see reference 2). Velocity distribution on the surface for several stream Mach numbers M, and extent of supersonic flow for $M=0.83$.

potential $\phi(x, y)$ is

$$\frac{\partial^2\phi}{\partial y^2}=\frac{\gamma+1}{a^*}\frac{\partial\phi}{\partial x}\frac{\partial^2\phi}{\partial x^2}, \tag{3}$$

where γ denotes the isentropic exponent. This is typically non-linear and exhibits the characteristic change-over from elliptic to hyperbolic type at the speed of sound. The discovery that underlies this theory is that, for a slender body of thickness ratio τ, the partial derivatives have the following relative orders of magnitude:

$$\left.\begin{aligned}&\partial(\ )/\partial x=O(1),\quad \partial(\ )/\partial y=O(\tau^{\frac{1}{3}}),\\ \text{while}\qquad &\phi=O(\tau^{\frac{2}{3}}).\end{aligned}\right\} \tag{4}$$

Moreover, both von Kármán and Guderley pointed out the similarity rule that is implied by these results, and which gives the engineer badly needed assistance in filling in the gap in his knowledge around the speed of sound. At the same time, Eq. (3) gives the theoretician a relatively simple equation with which to study the essential features of transonic flow. Certainly its potentialities have hardly been exploited to date.

Finally, there is one more device for the study of these inherently non-linear flow equations, at least in two-dimensional cases; and that is to transform them to the hodograph plane, i.e., to introduce a change of variables by which the flow velocity components u and v, say, become the independent variables, rather than x and y. The equations then become linear, which is not completely surprising because the essential non-linearity arose from the fact of the density's being a function of the speed, $(u^2+v^2)^{\frac{1}{2}}$. This method of linearization was discovered by Molenbroek[7] in 1890, and was applied to this problem in 1904 by the Russian, Chaplygin,[8] working on the theory of gas jets. It permits the construction of solutions by superposition of elementary solutions—in this case certain hypergeometric functions. The essential difficulty, unfortunately, then resides in the question of satisfying boundary conditions. Naturally, one does not in most cases have any way to satisfy the conditions for flow around a given solid contour in terms of boundary conditions established in the hodograph plane.

Nevertheless, the method has the unique advantage of supplying exact solutions of the equations of steady plane potential flow of a perfect gas. Thus its importance

[7] P. Molenbroek, "Über einige Bewegungen eines Gases bei Annahme eines Geschwindigkeitspotentials," Archiv d. Mat. Phys. 9, 157–195 (1890).

[8] S. A. Chaplygin, "On gas jets," Scientific Annals of the Imperial University of Moscow, Physico-mathematical Division, No. 21, Moscow, 1904 (translated in N.A.C.A. Tech. Memo. No. 1063 (1944)).

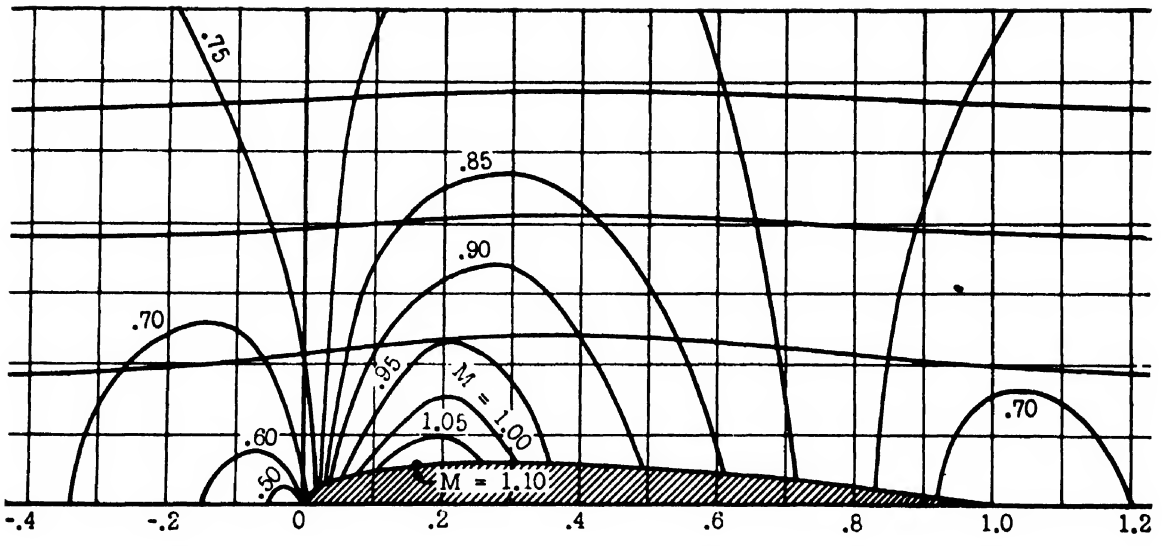


FIG. 4. Transonic potential flow over an airfoil, according to calculations of Emmons (see reference 4). Streamlines and lines of constant Mach number.

in research, and as a test for approximate methods, is great. Being practically unable to satisfy the boundary conditions of given shapes, one naturally tries the next best thing, which is to find a likely solution, either by trial or by relating it to incompressible flow in the proper limiting case, and then to see what sort of physical situation it gives. There are a number of different mathematical techniques for carrying out this sort of calculation, and unfortunately they all seem to be pretty complicated. The hodograph transformation is practically always multiple-sheeted with branch points that cause great difficulties in series expansions. Consequently, in many of these investigations the physics seems to be buried under tremendous amounts of analysis, and there is also an apparently endless controversy, disconcerting to the outsider, among various students of the subject, who fail to find much of value in each other's methods.

2. THEORETICAL MIXED FLOWS

Nevertheless, the hodograph method, in common with all the other, approximate, methods that have been applied to the transonic flow problem, has yielded one result of startling significance to the aeronautical engineer. It produces isentropic, potential flows that involve local acceleration through the velocity of sound, an imbedded region of supersonic flow near a body, and deceleration back to subsonic speeds. This is intriguing because it contrasts so radically with the situation commonly observed in nature—where the transition from supersonic to subsonic flow is accomplished by means of a violent, non-isentropic process, namely, a shock wave.

Apparently the first example of this theoretical phenomenon was provided by Taylor,[1] who found it in a symmetrical nozzle, using the series expansion mentioned above. One of his results is reproduced in Fig. 1. Analogous results were obtained by Görtler[3] (Fig. 2) and Kaplan[2] (Fig. 3), both using the method of expansion in powers of the thickness, and by Emmons[4] in his purely numerical solutions (Fig. 4).

The first exact solution of this type was found by Ringleb,[9] by means of the hodograph transformation. Ringleb started with a very simple solution of the hodograph equations, being, in fact an almost obvious generalization of the incompressible case of flow around a 180° corner. His result is reproduced in Fig. 5. It is not only of interest as the first exact mixed solution, but it presents two other surprising features. The first of these is the great magnitude of the supersonic Mach numbers attained in the flow (about 2.5), and the second is the appearance of singularities, called "limit lines" which seem to invalidate the solution when too great wall curvature is attempted. In other words, any two streamlines such as (a) and (b) in Fig. 5 might form the walls of a channel, and according to the theory, this channel would permit isentropic acceleration to, and deceleration from, supersonic speeds. But streamlines between (c) and (d) cannot be traced beyond the limit line; according to the transformation, in fact, they double back on themselves there and cross other streamlines—in other words, they go over onto a different Riemann sheet.

When Ringleb's paper first appeared, in 1940, there was considerable speculation about the physical meaning of the limit line, and there was a natural tendency to identify them with shock waves. More recently, however, aerodynamicists have taken a more cautious

[9] F. Ringleb, "Exakte Lösungen der Differentialgleichungen einer adiabatischen Gasströmung," Zeits. f. angew. Mat. Mech. 20, 185–198 (1940). Abstract in J. Roy. Aero. Soc. 46, 403–404 (1942).

position on this question—although it remains controversial. In Ringleb's case, for example, the appearance of the limit line at streamline (c) is accompanied by infinite curvature of that streamline. There is considerable doubt as to whether it could ever appear, anywhere in the flow, in the case of flow over a smooth solid body of fixed shape. This question has been studied by Tollmien,[10] Schaefer,[11] Nikolskii and Taganov,[12] and Friedrichs,[13] not all of whom have arrived at the same conclusions. However, it is not the principal concern of this paper, and we shall leave it at this point.

The resemblance of Ringleb's flow pattern to the flow over the top of an airfoil at high subsonic speed inspired Tsien and Kuo[14] to attack problems of this type with the hodograph method; i.e., flow of a parallel, subsonic stream past a cylindrical obstacle. One of their results is shown in Fig. 6. Here we have a symmetrical, nearly elliptic cylinder, in plane potential flow, and the Mach numbers in its imbedded supersonic zone reach a maximum value of 1.24.

It should be clear at this point that perfect-fluid theories, both exact and approximate, predict smooth potential mixed flows, in channels and past obstacles, that involve both acceleration and deceleration through the speed of sound. It is obviously of great practical and scientific importance to explore thoroughly the prospects of achieving this type of flow experimentally. The Tsien-Kuo cylinder clearly produces no drag, in the absence of skin friction, while the shock wave, if one occurred at a Mach number near 1.24, would be of strength sufficient to involve considerable drag, and moreover would probably cause boundary-layer separation and all of the undesirable consequences thereof. There would be great merit in learning how to reproduce this smooth mixed flow in nature.

The absence, to date, of any convincing experimental confirmation of this type of flow, however, is impressive. Invariably, when the local supersonic speeds of a transonic situation are great enough to make the determination reliable, we find that the supersonic region terminates in a shock wave, or in a series of shock waves, rather than in a smooth deceleration through the speed of sound.

3. POSSIBLE EXPLANATIONS

What is the reason for this glaring discrepancy between theory and experiment? Three different plausible explanations have been suggested: (a) that the theory is in error by the neglection of viscosity; (b) that the potential flows exist for only special shapes and flow speeds and do not have neighboring solutions. Hence they cannot be actually produced in experiments; (c) that the potential flows are unstable to small disturbances in the airstream.

In support of the first of these, Professor Liepmann has pointed out[15] that the smooth deceleration can be thought of as the result of effects carried rearward by the family of Mach waves originating farther forward, as sketched in Fig. 7(a), and that a distortion of the contour, as might be caused by a boundary layer, would suffice to cause an envelope of waves, as in Fig. 7(b). Such a discontinuity, of course, marks the breakdown of potential flow, and the pattern must be replaced by one involving a shock. But this would imply that a slightly different body, producing the Mach waves of Fig. 7(a) outside of its boundary layer, would permit smooth

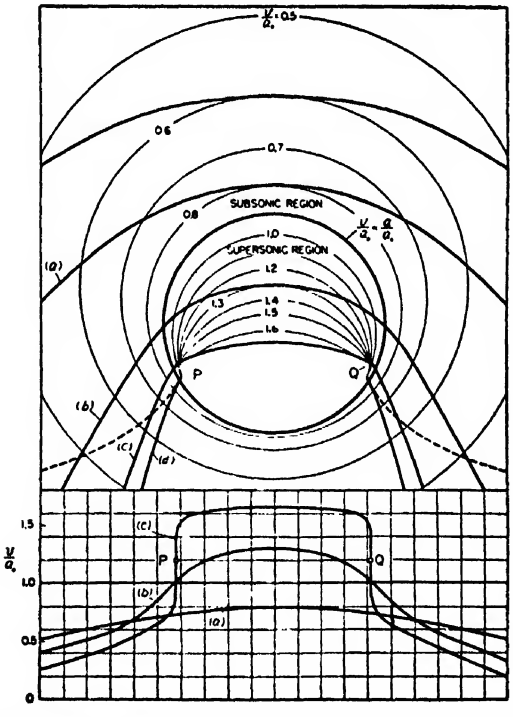


FIG. 5. Transonic flow around an obstacle, according to calculations of Ringleb (see reference 9). Top: streamlines and lines of constant speed. Bottom: velocity distributions along three streamlines.

[10] W. Tollmien, Grenzlinien adiabatischer Potentialströmungen. Zeits. f. angew. Mat. Mech. 21, 140–152 (1941).

[11] M. Schaefer, Enveloppenbildung krummliniger Machscher Wellen bei Strömung längs einer konvexen Wand, Technische Hochschule Dresden, 1944 [translated in U.S.A.F. A.M.C. Tech. Report No. F-TS-1203-1A (GDAM A9-T-13)]. Also, Auftreten eines Verdichtungsstosses in der Nachbarschaft einer konvexen singularitätenfreien Wandkontur. Technische Hochschule Dresden, 1944 [translated in U.S.A.F. A.M.C. Tech. Report No. F-TS-1206-1A (GDAM A9-T-14)].

[12] A. A. Nikolskii and G. I. Taganov, "Gas motion in a local supersonic region and conditions of potential-flow breakdown," Prikladnaya Matematika i Mekhanika 10, 481–502 (1946). [Translated in N.A.C.A. Tech. Memo. No. 1213 (1949).]

[13] K. O. Friedrichs, "On the non-occurrence of a limiting line in transonic flow," Inst. for Math. and Mech., New York University, Rep. IMM-NYU 165 (1947).

[14] H. S. Tsien and Y. H. Kuo, "Two-dimensional irrotational mixed subsonic and supersonic flow of a compressible fluid and the upper critical Mach number," N.A.C.A. T.N. 995 (1946). Also, Y. H. Kuo, "Two-dimensional irrotational transonic flows of a compressible fluid," N.A.C.A. T.N. 1445 (1948).

[15] H. W. Liepmann, "Investigations of the interaction of boundary layer and shock waves in transonic flow," Final Report, Contract W33-038 ac-1717 (11592), supplemental agreement No. 4 (S-4843), U.S.A.F. A.M.C., Wright Field, Dayton, Ohio.

potential flow. This proposal hardly seems adequate to explain the complete absence of flows of this type in experiments.

Moreover, if Liepmann's explanation is correct, it should be possible to achieve shockless mixed flows experimentally with the aid of boundary-layer removal. It seems important to decide whether or not the inviscid potential flow itself has any peculiarities that would make it impossible.

The second proposal, that the flows are very special ones, with no neighboring flows, is supported by Guderley and Busemann. Their physical argument is illustrated by Fig. 8, which is taken from Busemann's recent paper.[16] Suppose there exists a smooth mixed flow, and that a slight change is made in the body contour as shown. Since the local flow is supersonic, the effects of this alteration, in the new steady-flow pattern, will be confined to an "alley" leading to the sonic curve, from which point it is reflected back to the body, and so forth, as sketched. But according to this process it finally affects the nearly-sonic flow toward the rear, and we know that the pressure, density, and speed increments will be magnified in such a region. Thus a small change of the boundary shape must produce a large change in the flow pattern; i.e., there is no neighboring solution. Substantially the same argument has been put forth by Frankl.[17]

All this is quite analogous to the case of decelerating channel flow. It is easy to draw up a channel for smooth deceleration of supersonic flow, through sonic, to subsonic; in fact, the channel is simply a Laval nozzle in reverse (Fig. 9). But the flow is unreal, because there is

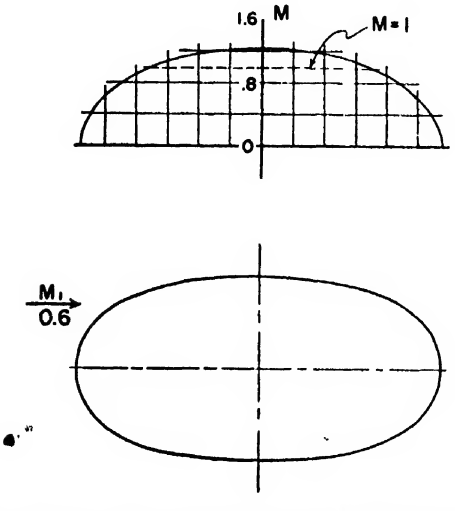


FIG. 6. Mach number distribution along a cylinder in transonic potential flow at a stream Mach number of 0.6, according to calculations of Tsien and Kuo (see reference 14).

[16] A. Busemann, "The drag problem at high subsonic speeds," J. Ae. Sci. 16, 337–344 (1949).

[17] F. I. Frankl, "On the formation of shock waves in subsonic flows with local supersonic velocities." Prikladnaya Matematica i Mekhanika 11, 199–202 (1947). (Translated in USAF–AMC Tech. Report No. F-TS-1223-IA, Dayton, Ohio, May, 1949.)

no neighboring solution. A slight change of the contour at A produces violent effects at B. The real flow is the one involving a shock wave.

Thus, Busemann's is certainly a very plausible explanation. It would be indisputable if the perturbation at the sonic curve did not actually affect the whole flow pattern through the subsonic flow, but of course it does. To carry along the channel analogy, the trouble at B (Fig. 9) can be avoided by a small alteration at C or C', or by an alteration upstream, as at D or D', which would cancel the effect at A. Busemann argues that in the case of flow past a body it is unlikely that the effect on the subsonic flow will be just what is required to neutralize the disturbance and prevent the large changes toward the rear. Perhaps it does seem unlikely, but what is unlikely is often true.

Guderley[6] has attempted to give a mathematical theory of this effect, showing that there are eigensolutions that would invariably be brought in by any alteration of the contour, and whose coefficients increase without limit as the sonic line is approached. Unfortunately, Guderley's calculation becomes so complicated, requiring further approximations and assumptions as it progresses, that its conclusions are not convincing. For example, these solutions have infinite amplitudes at both the upstream and downstream sonic points, and since we postulate smooth flow upstream, it is not clear why they should not all be rejected, removing the trouble at the downstream end as well. I suspect that the singularities are not real ones, but have arisen from Guderley's method of approximation, and the form of his results seems to confirm this.

There is also the problem of reconciling the Guderley-Busemann-Frankl theory with the Taylor, Emmons, and Kaplan cases mentioned above. None of these were calculated for a peculiar or special body shape or for any unique combination of body shape and Mach number. It seems preposterous to think that they are some kind of singular solutions, without neighbors. But these are all approximate solutions, and how much can one conclude about existence and uniqueness from investigations of this sort? Again it seems "unlikely" that these calculations would not show some trouble if they were being forced to fit contours not specially suited for this type of flow. For example, one might have thought that Emmons' numerical work would diverge. This point is not conclusive, however, for the effects may be small at the Mach numbers handled, and may not have been noticed. Moreover, returning to our channel analogy, we must admit that there an approximate method (one-dimensional gas dynamics) fails entirely to predict the non-existence of the flow of Fig. 9.

4. THE STABILITY OF MIXED FLOWS

In this situation, my colleague, Professor Kuo,[18] has undertaken to investigate the stability of mixed po-

[18] Y. H. Kuo, "On the stability of two-dimensional smooth transonic flows," paper delivered at the Annual Meeting of the

tential flow by the standard method of tracing the histories of time-dependent small disturbances. To a considerable extent his method follows that of Kantrowitz,[19] who carried out an approximate, but highly illuminating investigation of the stability of decelerating channel flow. Kantrowitz considered a channel flow going from supersonic to subsonic and calculated the behaviors of expansion and compression disturbances coming from downstream. He pointed out that the compression part of either a compression or expansion pulse steepens to form a weak shock (as Riemann showed for a uniform field) and the pulse assumes a characteristic triangular shape. Furthermore, he showed that the expansion pulses are unable to penetrate the supersonic region, and, in fact, are "consumed" by their own shock waves. Compression pulses, on the other hand, which have shock waves at their heads, move into the supersonic flow and pile up there, so that the flow finally undergoes violent changes.

Following Kantrowitz, Kuo considers the effects of upstream-moving disturbances of the compression-pulse type, and assumes that they have undergone the Riemann steepening process and have taken on the characteristic triangular shape. Since the sonic region is most critical, and since disturbances travel slowly in this region, he assumes that the pulses take on the character typical of sonic flow; i.e., their derivatives have the relative magnitudes specified by the Kármán-Guderley transonic similarity theory and stated in Eqs. (4). This leads to an accompanying assumption about the relative size of the time derivatives; it turns out that there is only one possibility here that gives a reasonable result for the non-stationary flow.

These assumptions are sufficient to permit a solution for small disturbances superimposed on a steady potential mixed flow of a parallel stream past a cylinder, at least for the region near either sonic boundary.

Let the velocity potential of the flow be

$$\varphi'(x, y, t)=\Phi(x, y)+\phi(x, y, t), \tag{5}$$

where $\phi(x, y, t)$ denotes the potential of the small disturbance superimposed on the steady transonic flow $\Phi(x, y)$. Then, by virtue of the order-of-magnitude argument for the sonic region, the differential equation for $\phi(x, y, t)$ is reduced to the following approximate one:

$$(\gamma+1)(\Phi_x^{(1)}+\phi_x)\phi_{xx}-a^*\phi_{yy} +2\phi_{xt}+(\gamma+1)\Phi_{xx}^{(1)}\phi_x=0, \tag{6}$$

where $\Phi^{(1)}$ denotes $\Phi-a^*x$. Kuo approximates to $\Phi^{(1)}$ in linear form; i.e.,

$$\Phi^{(1)}=\alpha\xi+\beta\eta, \tag{7}$$

where α and β are constants, while ξ and η denote

FIG. 7. Sketch showing effect of boundary layer on transonic potential flow: (a) without viscosity; (b) with viscosity.

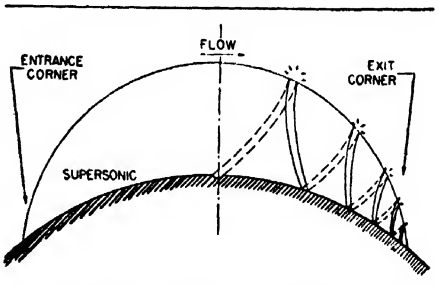


FIG. 8. Sketch showing effect on transonic potential flow of small alteration of boundary shape, according to Busemann (see reference 16).

distances measured parallel and perpendicular to the body contour from the sonic point. The boundary condition at the solid body is, similarly,

$$\phi_\eta=\phi_\xi(m+k\xi), \tag{8}$$

where m is the slope (relative to the free-stream direction) and k the curvature of the body surface at the sonic point.

Kuo finds the solution $\phi(x, y, t)$ in simple functions involving, besides initial conditions, the parameters

$$\left.\begin{aligned}&\nu=(\gamma+1)\alpha\\ \text{and}\qquad&\\ &\sigma=\tfrac{1}{2}(\nu+k/m).\end{aligned}\right\} \tag{9}$$

In the case of one-dimensional flow, his results reduce to those of Kantrowitz.[19]

For convex surfaces, $k<0$. At the accelerating sonic point, $\alpha>0$ and, presumably, $m>0$; this combination makes all the disturbances damp rapidly to zero, which confirms the observed stability of such flow. On the other hand, where $k<0$ and $\alpha<0$ (decelerating steady flow over a convex surface), the damping or growing of triangular pulses is influenced by the sign of m. When the decelerating sonic point occurs in a region of negative slope m, as is the case in Figs. 1–6, the disturbances increase in size in the sonic region, which indicates instability. But the theory brings forth an intriguing new idea: if m were positive here, the flow might again be stable. For example, if the imbedded supersonic area near an airfoil could be made to lie entirely *forward* of the maximum ordinate of the surface (Fig. 10), there would be a great reduction in the rate of growth of pulses as they pass through the sonic area, so that much larger disturbances could be tolerated; this might be termed pseudostability. Moreover, certain combinations of α, β,

Institute of Aero. Sciences, New York, January 24, 1950 (to be published).

[19] A. R. Kantrowitz, "The formation and stability of normal shock waves in channel flows," N.A.C.A. T.N. 1225 (1948).

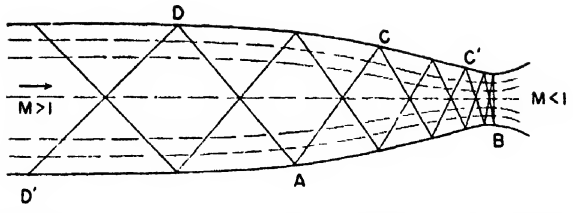


FIG. 9. Sketch showing reversed Laval nozzle for deceleration of supersonic flow to subsonic.

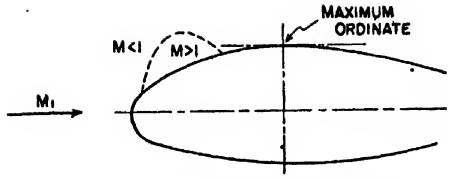


FIG. 10. Sketch showing possible stable transonic potential flow according to Kuo.

and m would result in complete damping of the pulses, i.e., stability.

So far as we know, there are no definitive measurements for flows of this type. Apparently it is possible to find airfoil shapes that have their supersonic regions entirely forward of their maximum ordinates. We are not yet sure that the peculiar combinations of α, β, and m mentioned above, for which the pulses are completely damped out, do occur. As Kuo has pointed out, there is a remarkable coincidence between his theoretical prediction and some recent experimental results of the N.A.C.A.[20] Entirely independently of Kuo's work, Nitzberg and Crandall at Ames Laboratory arrived at the empirical conclusion that the relative location of the sonic area and the "crest" of the airfoil is of significance in predicting the drag-divergence Mach number, M_D. To attribute this result to Kuo's stability effects is a long step, implying as it would that a shockless potential mixed flow exists before drag divergence or at least that the growth of small disturbances calculated by Kuo is the cause of drag divergence. A more conservative explanation would be that, although there are stationary shock waves, rather than potential flow, both below and above M_D, the shock waves do not cause separation of the boundary layer until $M=M_D$. Nevertheless, the coincidence is striking and, as Kuo has stated, seems to justify further research.

[20] G. E. Nitzberg and S. Crandall, "A study of flow changes associated with airfoil section drag rise at supercritical speeds," N.A.C.A. T.N. 1813 (1949).

5. CONCLUSION

In conclusion, it may be stated with assurance, that progress is now being made on this important question of shockless transonic flows. It might be pointed out explicitly that temporal instability and non-existence of neighboring solutions may represent the same phenomenon. One can explain the instability of an Euler column, for example, in either way, and Kantrowitz's results show that this is also true for the decelerating channel. Nevertheless, Guderley's, Frankl's, and Busemann's arguments have not suggested any stable decelerating cases, as Kuo's criterion suggests.

It seems advisable to undertake experimental confirmation of Kuo's results as soon as possible. We are studying this possibility at Cornell, but are realizing that the experimental difficulties may be considerable. One would like to have a really smooth potential flow to begin with, and to watch preselected disturbances travel into it. Unfortunately, there are usually an alarming number of disturbances naturally present, apparently originating in the boundary layer. We should like to suppress these, or to have so stable a flow that a disturbance larger than these would still be damped out. So far, we have not figured out how to accomplish this.

Analysis of Intensities Obtainable Using Pinhole and Slit Collimators in X-Ray Powder Diffraction Techniques

LEROY ALEXANDER
Department of Research in Chemical Physics, Mellon Institute, Pittsburgh, Pennsylvania
(Received February 6, 1950)

Other factors being equal, the integrated intensity of a powder diffraction reflection is a function of the geometry of the source, collimator, and sample. By means of a geometrical and algebraic analysis one can predict the relative intensities to be expected with pinhole or slit collimators of different dimensions. Important gains in intensity can be obtained by replacing pinholes with slits, and the advantage is about the same if the x-ray focal spot is viewed either longitudinally or laterally through slits of sufficient length. These results confirm experimental results reported by other investigators and also re-emphasize the potential value of the "side" windows of x-ray tubes, which are commonly neglected.

SEVERAL investigators[1–3] have recently pointed out the practicability of preparing Debye-Scherrer x-ray diffraction photographs by employing a pair of parallel slits rather than pinholes. The gain in intensity is great, being roughly proportional to the ratio of the slit length to the diameter of the pinholes employed in customary practice (for the case of the pinhole diameter equal to the slit width). The increased intensity is gained at the expense of decreased resolution and a shift in line positions, but both of these effects are very small for moderate slit lengths,[3] whereas the attendant intensity gain is several fold. Consequently such an arrangement is entirely satisfactory for the numerous applications of powder diffraction for which optimum line quality is not a prime consideration, and furthermore it permits a very substantial saving in exposure time and increase of x-ray tube life.

If the focal spot is viewed at some fixed small angle to the target face, say 6° or 7°, and if the two slits, the long dimension (if any) of the projected focal spot, and the sample axis are all parallel, it has been found[1,3] that for slits of sufficient length substantially equal gains in intensity are obtained whether the source is viewed (a) in a direction parallel to its long axis so that it appears foreshortened, the usual orientation when using pinhole techniques, or (b) in a direction normal to its long axis. In the present discussion orientations (a) and (b) will be respectively denoted the longitudinal and lateral arrangements. Thus by a relatively simple modification in camera design the "side" windows of an x-ray tube, commonly neglected by most diffractionists, can be put to good use. For this reason the recommendation has recently been made that x-ray tube manufacturers build four- rather than two-window diffraction tubes.[2]

The relative intensities to be expected with different collimator and source arrangements can be predicted with considerable accuracy by a geometrical and algebraic analysis. The mathematical development will be simplified by making the following initial assumptions: (a) The x-ray focal spot is a rectangle of uniform intensity. (b) The source-to-sample distance, x_3, is kept constant. If it were to be varied, an additional factor $1/x_3^2$ would need to be applied to each intensity equation. (c) The focal spot is viewed at some constant angle, α, with the target face. This requirement must be met since at angles less than about 8 or 10° the energy flux from the focus diminishes appreciably with decreasing angle.[4] (d) The x-ray beam impinges perpendicularly upon the sample, which takes the form of a flat sheet of negligible absorbing power for the x-rays. In actuality the results apply without serious error to a cylindrical Debye-Scherrer sample of low or moderate absorbing power. (e) The apertures S_1 and S_2 are rectangular or square in shape.

Let us set up a three-dimensional rectangular coordinate system with origin at the center of the x-ray tube focus as shown in Fig. 1. It will be noted that the widths of the slits S_1 and S_2 and the projected dimension of the focal spot as governed by the angle of view, α, are measured in the direction Z. The distances separating the source, slits, and sample are measured in the direction X; and the heights of the slits, the height of the area of sample irradiated, and the *non-projected* dimension of the x-ray focus are measured in the direction Y. For reasons that will become apparent as the analysis proceeds, one may consider the effects of the Z dimension as being virtually constant for all experimental

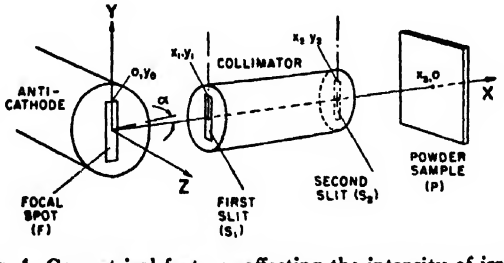


FIG. 1. Geometrical features affecting the intensity of irradiation of a powder diffraction sample. The case portrayed is that in which the focal spot is viewed laterally.

[1] A. Unmack and A. T. Jensen, J. Sci. Inst. Phys. Ind. 25, 399 (1948).
[2] A. Taylor, J. Sci. Inst. Phys. Ind. 26, 61 (1949).
[3] R. I. Garrod, J. Sci. Inst. Phys. Ind. 26, 162 (1949).
[4] A. Bouwers and P. Diepenhorst, Fortschr. a. d. Geb. d. Röntgenstr. 38, 894–898 (1928).

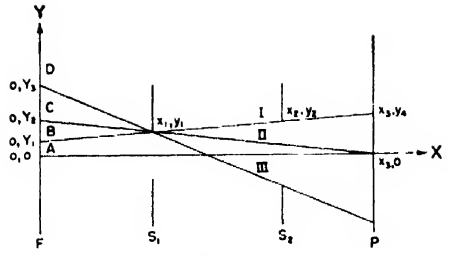


FIG. 2.

cases in which the apertures S_1 and S_2 have the same width (Z dimension) and in which at the same time the aperture width is either roughly equal to or larger than the projected (Z) dimension of the focal spot. These conditions obtain in the usual experimental cases, such as those to be compared in the present communication, and hence we may simplify the mathematical treatment by confining it to the XY plane.

Referring again to Fig. 1, the coordinates of the relevant features of our apparatus are: ends of focus, $(0, y_0)$ and $(0, -y_0)$; ends of slits 1 and 2, (x_1, y_1), $(x_1, -y_1)$ and (x_2, y_2), $(x_2, -y_2)$ respectively; intersection of x-ray beam axis and sample trace, $(x_3, 0)$. Four principal mathematical cases arise, depending upon the relative values of x_1, x_2, x_3, y_1, y_2, and y_0. Referring to Fig. 2, we may characterize these cases as A, B, C, and D according to whether y_0 falls in region A, B, C, or D respectively. Let Y_1, Y_2, and Y_3 be the respective zone boundaries AB, BC, and CD. These zone boundaries are the Y intercepts of the lines I, II, and III, and by consideration of the diagram their equations are seen to be

$$Y_1 = y_1 - \left(\frac{y_2 - y_1}{x_2 - x_1}\right)x_1 = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}, \tag{1}$$

$$Y_2 = y_1 + \left(\frac{y_1}{x_3 - x_1}\right)x_1 = \frac{x_3 y_1}{x_3 - x_1}, \tag{2}$$

$$Y_3 = y_1 + \left(\frac{y_1 + y_2}{x_2 - x_1}\right)x_1 = \frac{x_1 y_2 + x_2 y_1}{x_2 - x_1}. \tag{3}$$

The four principal mathematical cases may then be expressed as follows:

Case A. $Y_1 \geq y_0$.
Case B. $Y_2 \geq y_0 \geq Y_1$.
Case C. $Y_3 \geq y_0 \geq Y_2$.
Case D. $y_0 \geq Y_3$.

Further geometrical cases arise when the apertures S_1 and S_2 differ very greatly in length. Thus when $y_1 \ll y_2$, Y_1 becomes negative, and when $y_1 \gg y_2$, y_4 becomes negative (see Fig. 3). However, these unusual arrangements have little practical value, and it is desirable to avoid complicating the present discussion by introducing these additional geometrical cases.

Before carrying out the mathematical analyses it will be useful to write the equations for the intersections of five lines with the line x_3 (see Fig. 3):

$$y_3 = y_0 + (y_1 - y_0)\frac{x_3}{x_1} = y_1 + (y_1 - y_0)\left(\frac{x_3 - x_1}{x_1}\right), \tag{4}$$

$$y_4 = y_1 + (y_2 - y_1)\left(\frac{x_3 - x_1}{x_2 - x_1}\right)$$

$$= y_2 + (y_2 - y_1)\left(\frac{x_3 - x_2}{x_2 - x_1}\right), \tag{5}$$

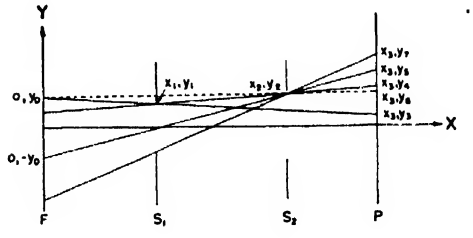


FIG. 3.

$$y_5 = -y_0 + (y_0 + y_2)\frac{x_3}{x_2} = y_2 + (y_0 + y_2)\left(\frac{x_3 - x_2}{x_2}\right), \tag{6}$$

$$y_6 = y_0 - (y_0 - y_2)\frac{x_3}{x_2} = y_2 - (y_0 - y_2)\left(\frac{x_3 - x_2}{x_2}\right), \tag{7}$$

$$y_7 = -y_1 + (y_1 + y_2)\left(\frac{x_3 - x_1}{x_2 - x_1}\right)$$

$$= y_2 + (y_1 + y_2)\left(\frac{x_3 - x_2}{x_2 - x_1}\right). \tag{8}$$

We now proceed to analyze the rather typical case B. Figure 4 shows the geometrical properties of this case. At the right of the sample trace, P, the intensity of illumination of the sample as a function of y is depicted by the hatched contour. The relative integrated intensity of illumination may in general be represented by

$$I = \int_{-\infty}^{+\infty} I_y dy,$$

which becomes because of symmetry

$$I = 2\int_0^{\infty} I_y dy. \tag{9}$$

For the present case the function I_y is different for each of the regions $0 - y_3$, $y_3 - y_4$, and $y_4 - y_5$, so the integral for convenience may be broken up into three terms:

$$I = 2\int_0^{y_3} I_y dy + 2\int_{y_3}^{y_4} I_y' dy + 2\int_{y_4}^{y_5} I_y'' dy.$$

These three integrals are seen by inspection of Fig. 4 to be given respectively by the areas of a rectangle, trapezoid, and triangle, so that the expression for I becomes

$$I=2[y_3I_3+\tfrac{1}{2}(y_4-y_3)(I_3+I_4)+\tfrac{1}{2}(y_5-y_4)I_4].$$

At any point in the zone $-y_3$ to $+y_3$ the entire focal spot can be seen and, hence, the intensity of irradiation is a maximum, I_m. Substituting I_m for I_3 and combining terms, we obtain

$$I=I_m(y_3+y_4)-I_4(y_3-y_5) \quad \text{(case B)} \qquad (10)$$

in which y_3, y_4, and y_5 are given by Eqs. (4), (5), and (6) respectively. Now the intensity I_4 is equal to I_m times the fraction of the focal spot visible at the point x_3, y_4. By comparing Figs. 2 and 4 this is seen to give

$$I_4=I_m\cdot\frac{y_0+Y_1}{2y_0},$$

$$=(1/2)I_m\left[\frac{x_2y_1-x_1y_2}{y_0(x_2-x_1)}+1\right]. \qquad (11)$$

Substitution of Eqs. (4)–(6) and (11) in (10) leads to a somewhat complicated algebraic expression for the integrated intensity of illumination of the sample powder as a function of the dimensions and positions of the collimating apertures and the dimensions of the focal spot. Because of the complexity of the fully substituted Eq. (10) it is more convenient to evaluate it for specific experimental cases by first evaluating y_3, y_4, y_5, and I_4 and then substituting the numerical values in Eq. (10).

It would be tedious and unnecessary to reproduce in full here the complete analyses for cases A, C, and D. Suffice it to say that by a procedure very similar to that described above for case B the appropriate integrated intensity expressions can be derived. The expressions for all four cases are summarized herewith:

Case A.

$$I=2I_m\frac{x_3}{x_2}y_2. \qquad (12)$$

Case B.

$$I=I_m(y_3+y_4)-I_4(y_3-y_5), \qquad (10)$$

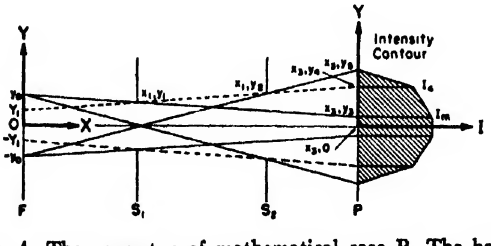


FIG. 4. The geometry of mathematical case B. The hatched contour depicts the intensity of sample irradiation as a function of the coordinate y.

TABLE I. Values of the x and y parameters for four typical experimental situations.

Parameter	Values of the parameters in mm (1)	(2)	(3)	(4)
x_1	60	60	60	60
x_2	100	100	100	100
x_3	150	150	150	150
y_0	0.500	4.000	0.500	4.000
y_1	0.375	0.375	2.750	2.750
y_2	0.375	0.375	2.750	2.750

TABLE II. Quantities computed in the evaluation of the integrated intensities for the four experimental situations of Table I.

Quantity	(1)	(2)	(3)	(4)
Y_1	0.375	0.375	2.750	2.750
Y_2	0.625	0.625	4.580	4.580
Y_3	1.500	1.500	11.000	11.000
Mathematical case	B	D	A	B
y_3	0.187	—	—	0.870
y_4	0.375	—	—	2.75
y_5	0.812	—	—	6.13
I_4	$0.875I_m$	—	—	$0.844I_m$
Calculated intensity in I_m units	1.11	0.26	8.25	8.06
Relative intensity (rectangular apertures)	13.4	3.2	100.0	97.5
Relative intensity (circular pinholes)	10.5	2.5	100.0	97.5

where

$$I_4=(1/2)I_m\left[\frac{x_2y_1-x_1y_2}{y_0(x_2-x_1)}+1\right]. \qquad (13)$$

Case C.

$$I=I_3(y_4-y_3)+I_4(y_3+y_5), \qquad (14)$$

where

$$I_3=I_m\frac{y_1}{y_0}\left(\frac{x_3}{x_3-x_1}\right) \qquad (15)$$

and I_4 is given by Eq. (13).

Case D.

$$I=2I_m\frac{x_3y_1y_2}{y_0(x_2-x_1)}. \qquad (16)$$

Equations (7) and (8), although not explicitly involved in the final intensity formulas, are essential to the derivation of Eqs. (12) and (16), respectively.

We may now return briefly to the observation made earlier that the Z dimensions of the apparatus could justifiably be disregarded in the present analysis. This action can be seen to be completely justified when the apertures S_1 and S_2 have the same width ($2z_1=2z_2$) and when at the same time the projected source dimension, $2z_0$, is equal to or less than $2z_1$, because this is simply mathematical case A applied to the Z dimension, and Eq. (12) shows I to be independent of y_0 (or z_0 in the present application). When z_0 begins to exceed z_1

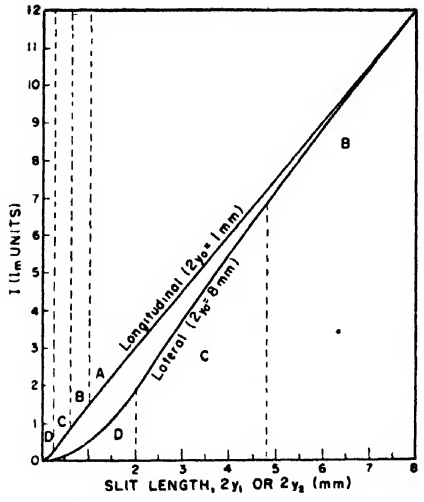


FIG. 5. Variation of integrated intensity, I, with slit length for a pair of identical rectangular slits of length $2y_1=2y_2$. The upper curve is for longitudinal and the lower for lateral view of the focus. Rectangular focal spot 1×8 mm; $\alpha=7°$; $x_1=60$, $x_2=100$, $x_3=150$ mm; slit width 0.75 mm. A, B, C, and D indicate applicable mathematical cases.

we are dealing with case B, but the application of Eq. (10) demonstrates that I diminishes very little until z_0 exceeds about $1.5z_1$.

For purposes of illustration we now apply these mathematical results to typical experimental conditions. Let $x_1=60$, $x_2=100$, and $x_3=150$ mm, and suppose the focal spot is a rectangle 1×8 mm in size. Furthermore, let us suppose that the focus is viewed at such an angle (7° in the present instance) that when observed in the longitudinal direction it appears as a square 1×1 mm in size. When viewed laterally, then, it will have the apparent dimensions 0.12×8 mm. Let us compare the integrated intensities of illumination for four important situations, including viewing the focal spot through: (1) two 0.75 mm pinholes longitudinally (customary arrangement using "end" windows), (2) two 0.75 mm pinholes laterally (using "side" windows), (3) two 0.75×5.5 mm slits longitudinally, (4) two 0.75×5.5 mm slits laterally. The proper values of the six x and y parameters for these four cases are given in Table I.

Table II lists the values of the various quantities which are calculated during the process of determining the integrated intensities for these four experimental situations. First, the quantities Y_1, Y_2, and Y_3 are computed, from which the appropriate mathematical treatment (case A, B, C, or D) is selected for each set of experimental conditions. Experimental cases (1) and (4) are found to fall in mathematical class B, and accordingly the quantities y_3, y_4, y_5, and I_4 are evaluated preliminary to the application of Eq. (10). Experimental cases (2) and (3) are found to fall in mathematical classes D and A respectively, for which no further quantities need be evaluated. Next the integrated intensity in I_m units is calculated using the appropriate equation, (10), (12), (14), or (16). The next row of data gives the relative intensities for the four cases referred to case (3) as 100. In the last row the relative intensities for the square apertures of edges $2y_1$ and $2y_2$ of cases (1) and (2) are converted to the equivalent values for circular apertures of diameters $2y_1$ and $2y_2$ by applying the factor $\pi r^2/(2r)^2=0.78$. The figures in the last row of Table II, then, give a direct comparison of the intensities to be expected when employing a pair of identical pinholes or a pair of identical rectangular slits.

The results given in Table II demonstrate clearly that a pair of pinholes is greatly inferior to a pair of slits of moderate length from the standpoint of irradiation intensity. Furthermore, the gain in intensity is found to be nearly the same whether the source be viewed longitudinally or laterally. These results agree with the experimental findings of Garrod[3] except for the precise extent of the intensity increases, which would hardly be expected to agree more closely because of differences in the respective experimental conditions. The second column of Table II shows a marked loss of intensity on viewing a rectangular focus laterally through a pair of pinholes with diameters much smaller than the length of the focus. This is, of course, the reason for the customary neglect of the "side" windows of the x-ray tube when only pinhole collimators are available.

The intensities provided by rectangular slits of various lengths are portrayed graphically in Fig. 5. The following typical conditions are assumed: $x_1=60$, $x_2=100$, $x_3=150$ mm; identical slits ($y_1=y_2$ and $z_1=z_2$); 1×8 mm rectangular focus viewed at 7°. The slit width ($2z_1$ or $2z_2$) is constant and equal to 0.75 mm. Figure 5 shows that the longitudinal direction of view is much preferable to the lateral for small values of $2y_1$ and $2y_2$, that is, when the slits are so short as to approach a square in shape. As the length of slit increases, the lateral intensity approaches the longitudinal, until when the slit length equals the length of the focus ($2y_1=2y_2=8$ mm in the present instance) the two techniques give equal intensities.

An analysis somewhat similar to the above has been carried out by Bolduan and Bear[5] for the special conditions inherent in low angle diffraction techniques. The results are applied to the design of collimators calculated to give optimum resolution and intensity.

[5] O. E. A. Bolduan and R. S. Bear, J. App. Phys. 20, 983–992 (1949).

Average Electron Density Measurements by Low Angle X-Ray Scattering*

H. M. Barton, Jr. and R. Brill
Polytechnic Institute of Brooklyn, Brooklyn, New York, and Phillips Petroleum Company, Bartlesville, Oklahoma
(Received March 8, 1950)

According to Guinier the low angle x-ray scattering of particles should disappear when they are immersed in a fluid of equal electron density. This furnishes a way of measuring the electron density of powders, providing a suitable impregnating fluid can be found. Experiments of this type have been carried out with carbon black. A method of calculating the low angle scattering coefficient defined by Warren is given for impregnated samples. Measurements of the total low angle scattering intensity were made using a double crystal spectrometer and Geiger counter. No liquid was found that would give complete matching but approximate matching was obtained with phosphoric acid solution. The data extrapolate to show an electron density of $1.1N \pm 0.1N$ where N = Avogadro's number. This compares with a value of $0.99N$ as calculated from the helium immersion mass density measurement on the same sample.

I. INTRODUCTION

IT has been stated by Guinier[1] that if particles of average electron density† ρ_e are suspended in a medium of electron density ρ_e', they will give low angle scattering as if they had electron density $(\rho_e - \rho_e')$ and were in a vacuum. This, of course, means that if there is no difference in electron density between the particles and their surroundings there will be no low angle scattering. This is analogous to the optical effect that particles immersed in a fluid of the same index do not scatter light.

Kratky, Secora, and Treer[2] investigated this experimentally for cellulose, finding in a qualitative way that the intensity of low angle scattering went to a low value when the particles were immersed in ethyl iodide, and that there was more intensity with fluids having greater or less electron density than the ethyl iodide.

Kratky and Wurster[3] extended the above experiments, measuring quantitatively the effect of different impregnating fluids on the small angle scattering of cellulose. They concluded that the intensity is proportional to the cube of the difference in electron densities. Their measurements were carried out using the photographic method. This procedure is rather complicated and requires many corrections, the magnitude of which may often be many times the measured value.

We used a more direct method, given by Warren.[4] By this method the total intensity scattered in the low angle region is measured by means of a double crystal spectrometer fixed at the maximum intensity angle. Warren defined a coefficient μ_{SA} which describes the reduction in primary beam intensity by a unit mass per unit area due to low angle scattering, in addition to that by the mass absorption; and showed it to be proportional to the integrated energy in low angle scattering. He derived the relation:

$$\mu_{SA} = \mu_m \frac{\ln(I_B/I_A)}{\ln(I_0/I_B)}, \tag{1}$$

where μ_m is the mass absorption coefficient for powders, I_A, I_B, and I_0 are the relative intensities measured with the sample in position A between the crystals, in position B in front of the Geiger counter, and out of the beam, respectively. The particles were in air.

II. SCATTERING COEFFICIENTS FOR IMPREGNATED SAMPLES

For samples impregnated with an ordinary liquid of mass absorption coefficient μ_i, we may set up the equation for the intensity at position A between the crystals

$$I_A = I_0 \exp[-\mu_i m_i - (\mu_p + \mu_{SA}) m_p], \tag{2}$$

where m_i is the mass per unit area of the impregnating liquid in the beam, μ_p is the mass absorption coefficient for powders of the particles, and m_p is the mass per unit area of the particles.

In position B in front of the Geiger counter where the low angle radiation is also registered:

$$I_B = I_0 \exp[-\mu_i m_i - \mu_p m_p], \tag{3}$$

and also we have for the total mass per unit area:

$$m_t = m_i + m_p. \tag{4}$$

Combining these three equations we get

$$\mu_{SA} = \frac{(\mu_p - \mu_i)\ln(I_B/I_A)}{\ln(I_0/I_B) - \mu_i m_t}, \tag{5}$$

which reduces to the Warren equation when μ_i is negligible. In the case where $\mu_p = \mu_i$, Eq. (5) becomes indeterminate, but this indeterminancy may be re-

* This is a condensation of a thesis presented by H. M. Barton in partial fulfillment of the requirements for a Masters degree at

[1] A. Guinier, J. chim. phys. 40, 133 (1943).

† $\rho_e = (n/M)\rho N$ where n is the average atomic number, M is the average atomic weight, ρ is the mass density, and N is Avogadro's number.

[2] Kratky, Secora, and Treer, Zeits. f. Elektrochemie 48, 587 (1942).

[3] O. Kratky and A. Wurster, Zeits. f. Elektrochemie 50, 249 (1944).

[4] B. E. Warren, J. App. Phys. 20, 96 (1949).

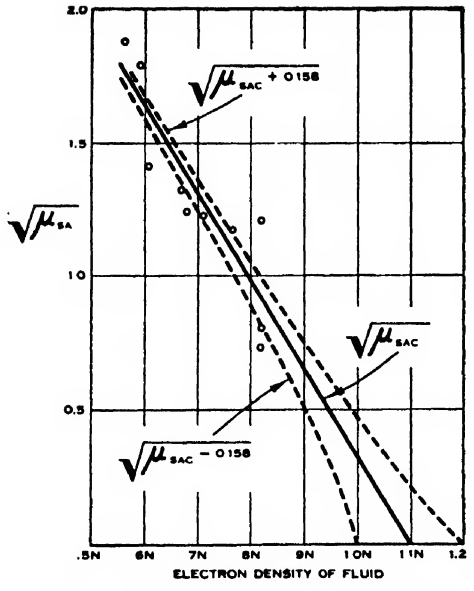


FIG. 1. Average electron density of impregnating fluid *vs.* square root of low angle scattering coefficient. μ_{SAC} is least squares calculation.

moved, giving

$$\lim_{\mu_p \to \mu_i} \mu_{SA} = \frac{\mu_i \ln(I_B/I_A)}{\ln(I_0/I_B) - \mu_i m_i}. \tag{6}$$

This case will seldom be encountered, however.

III. RELATION OF SCATTERING TO RELATIVE ELECTRON DENSITY

The intensity of low angle scattering is generally considered to be proportional to the square of the number of electrons per particle, when the medium may be neglected. If the medium has appreciable electron density and if Guinier's statement holds, then the particles would scatter as if the electron density of the particles were diminished by that of the medium, and the effective number of electrons per particle would be $(\rho_e - \rho_e')V$ where V is the volume of a particle. The low angle scattering is thus seen to be proportional to $(\rho_e - \rho_e')^2$, or

$$\mu_{SA} = b(\rho_e - \rho_e')^2. \tag{7}$$

If this equation holds, then it will not be necessary to actually match the electron densities of the particles and fluid, but only to define the parabola (7) with sufficient accuracy.

IV. EXPERIMENTAL

To test this, a double crystal spectrometer was set up using calcites in the parallel position. The half-width at half-maximum of crystal B was about 19 seconds. Copper $K\alpha_1$ radiation was used, and the resultant radiation was measured with Geiger counter equipment removed from a Norelco x-ray spectrometer.

Experiments were made on impregnated carbon black. The fluids finally used were solutions of phosphoric acid. It was not possible to match the electron density of the carbon black with these solutions, so it was necessary to extrapolate to find the match point. It is difficult to find liquids with high electron densities which do not also have very high absorption. This requires reducing the thickness of the sample, but the sample must be kept thick enough to give sufficient low angle scattering. Some salts such as magnesium nitrate have suitable electron densities and mass absorptions. It proved possible to make pastes with molten salts but on solidifying the salts gave their own low angle scattering. Construction of a hot holder was not undertaken.

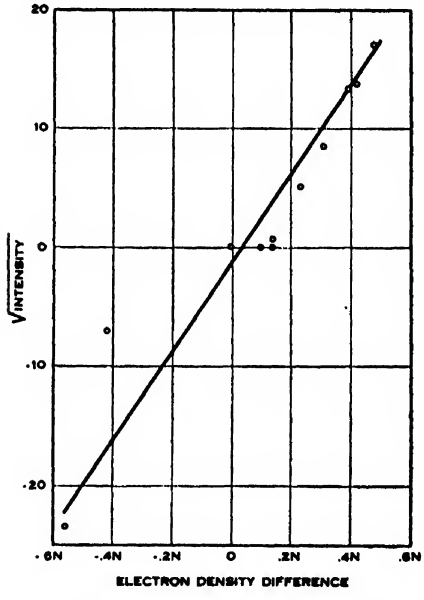


FIG. 2. Replot of data by Kratky and Wurster. Electron density difference between impregnating fluid and calculated electron density difference of hydrocellulose *vs.* square root of low angle scattering intensity. Intensity square roots are plotted as negative values for negative values of electron density difference.

V. RESULTS

Figure 1 shows the data plotted as the square root of μ_{SA} *vs.* the electron density of the fluid. The straight line is that obtained by least squares approximation to Eq. (7). It extrapolates to the axis at an electron density of $1.1N$. The average error expected for a measurement of μ_{SA} is:

$$p = \pm\left(\frac{\Sigma\Delta^2}{i(i-1)}\right)^{\frac{1}{2}}\left[1+\left(\frac{1}{2(i-1)}\right)^{\frac{1}{2}}\right], \tag{8}$$

where Δ is the difference between the least squares value and the experimentally measured value, and i is the number of measurements. For these data this comes out to be 0.158. The dashed curves were obtained by adding and subtracting this amount to the least squares curve and taking the square root. They are very close to

straight lines except near the intercept. The positive error curve does not actually go through the axis, but is shown extrapolated to get a rough idea of the error. These show an electron density for the carbon black of $1.1N \pm 0.1N$, corresponding to an electron density of $0.99N$ as calculated from the mass density determined by helium immersion of 1.98 g/cc.[5]

It would be highly desirable to obtain data closer to the point where the electron densities match. Experimentally this will require higher precision than was obtained in this work. It should be noted that the logarithm of the ratio of I_B to I_A occurs in Eq. (4). When there is little low angle scattering, this ratio is close to unity. Noting that when x is small the $\ln(1+x)$ is approximately x, it is easily seen that when the ratio is 1.01, a 1 percent error in measurement of the ratio will give a 100 percent error in the result. Measurements of intensities near the minimum of the parabola must therefore be made with great precision. Normally the extrapolation technique will give better results.

According to the paper by Kratky and Wurster,[3] the total intensity of small angle scattering should be proportional to $(\rho_e - \rho_e')^2$. Figure 2 shows their data replotted as the square root of the corrected intensities *vs.* $(\rho_e - \rho_e')$. We have the impression that a straight line fits with sufficient accuracy their results thus plotted to show the quadratic relation. Since it is not possible to obtain reasonable results with our data using the cubic assumption, we feel that the second power called for by theory is true.

We are indebted to Professor I. Fankuchen for several helpful discussions during the course of this work.

[5] L. R. Sperberg and H. M. Barton, Rubber Age 63, No. 1, 1 (April, 1948).

A Removable Intermediate Lens for Extending the Magnification Range of an Electron Microscope

J. Hillier
RCA Laboratories, Princeton, New Jersey
(Received February 27, 1950)

A removable intermediate lens is described which extends the range of magnification of a conventional instrument to 25:1 without sacrificing the accessibility of the objective and projection lens polepieces. The distortion, sensitivity to changes in accelerating potential and reproducibility of calibration are studied and found to be equal to, or better than in the standard arrangement. The system provides two values of the magnification which are essentially independent of the accelerating potential.

I. INTRODUCTION

In conventional electron microscopes it has been customary to use an objective of fixed magnification in the neighborhood of 100× and single projection lenses of variable magnification with maximum values in the range 200 to 300. In other words the maximum electronic magnification by the entire instrument is usually in the range 20,000 to 30,000 diameters. Values greater than this are possible but have been used only occasionally owing to a number of technical difficulties including in particular, the high bombardment of the specimen necessary to provide sufficient image intensity. While the projection lenses are usually variable, more than a fourfold variation is rarely used owing to the introduction, at the lower values, of excessive image distortion and to an excessive limitation of the field of view. The minimum magnification attainable in the example cited above is in the range 5000 to 7500 diameters. However, these values are not sufficiently low for many uses of the electron microscope. For survey purposes or for correlation with the light microscope, magnifications in the range 500 to 1500 are desirable. It has been well known by those familiar with electron optics that the range of magnification and the field of view can be extended by the use of multiple projection lenses. Hillier[1] and Prebus[2] used double projection lenses each of which was weaker than a single conventional lens but together gave both a wide range of magnification (including values higher than those attainable with a single lens) and an increased field of view. This system did not find favor at the time because the lenses were not easily accessible and for some parts of the range excessive distortions were encountered. Marton suggested that lower magnifications might be included in the range by interposing a weak and variable lens between the objective and projector of a conventional instrument. More recently le Poole[3] has built an instrument in which he combined these two ideas introducing two lenses between the objective and projector. By properly activating one or two of these three projection lenses he is able to achieve any magnification he desires between zero and a maximum in the range 50,000 to 100,000 diameters. The adjustment can be made so that there is minimum distortion and satisfactory field of view at all magnifications.

In the design of electron microscopes easily accessible

[1] J. Hillier, U. S. Patent 2,323,328.
[2] A. F. Prebus, "Electron Microscopy," *Colloid Chemistry* (Reinhold Publishing Corporation, New York, 1944), edited by Jerome Alexander, Vol. V.
[3] J. le Poole, Philips Tech. Rev. 9, 33, 1947.

TABLE I.

Tap no.	Total shunt resistance ohms	Instrument magnification at approximately 50 kv
1	400	7700
2	490	9000
3	580	10500
4	715	13000
5	880	14000
6	1060	15500
7	1330	17000
8	1720	18500
9	2720	20000
10	inf.	23400

polepieces have been considered of prime importance. This arises through the necessity of keeping the polepieces scrupulously clean. It has had the further advantage that a wide range of magnification could be obtained by the easy substitution of different lens combinations. For most work this has been found to be quite satisfactory. However, the convenience of a multiple lens system which enables the operator to survey his specimen at low magnification without breaking the vacuum, is not one to be overlooked.

In order to obtain the convenience of multiple lens operation without sacrificing polepiece accessibility in the electron microscope, a removable intermediate lens was designed which provides a range of magnification of 25:1 without changing polepieces. This new lens also reduces the distortion in the system, requires only a slight change in the electrical circuit of the RCA Model EMU, and provides the range of magnification in ten reproducible steps. The lens described was designed specifically for the standard RCA Model EMU in use in this laboratory.

II. DESCRIPTION OF REMOVABLE INTERMEDIATE LENS

Figures 1 and 2 are photographs of the complete unit. It is simply a lens spool consisting of two Armco iron end plates (one of which carries the central tube) threaded into a brass spacer which also acts as the lens-gap. The vacuum seal is achieved in the usual way by soft-soldering the threaded joints. The outer faces of the end plates are provided with projections which are machined respectively identical to the top and bottom ends of the intermediate chamber casting which it replaces in the instrument. The spool is surrounded by the usual sheath of Armco iron to complete the magnetic circuit. The spool is wound with insulated wire and must have a minimum resistance of 400 ohms. In our experimental model 9000 turns were wound on the coil which then had a resistance of 450 ohms. Connections to the coil were brought out into a shielded cable long enough to reach from the intermediate image position to the magnification control on the panel.

III. THEORY AND PERFORMANCE

The projection lens power supply in the Model EMU consists of a constant current regulator supplying 150 ma to the lens circuit and several regulator tube filaments. Variations in magnification are obtained by reducing the current in the projection lens; this being accomplished by a system of shunting resistors on a ten position step switch. (See Fig. 3) Table I gives the values

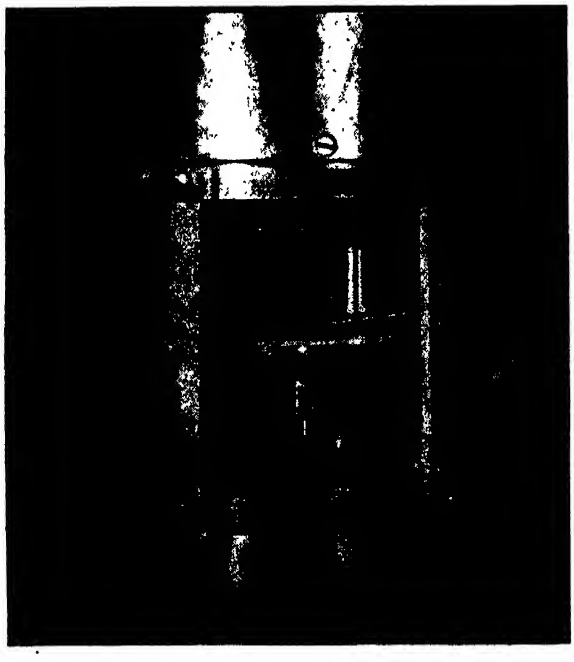


FIG. 1. A photograph showing the intermediate lens in position.

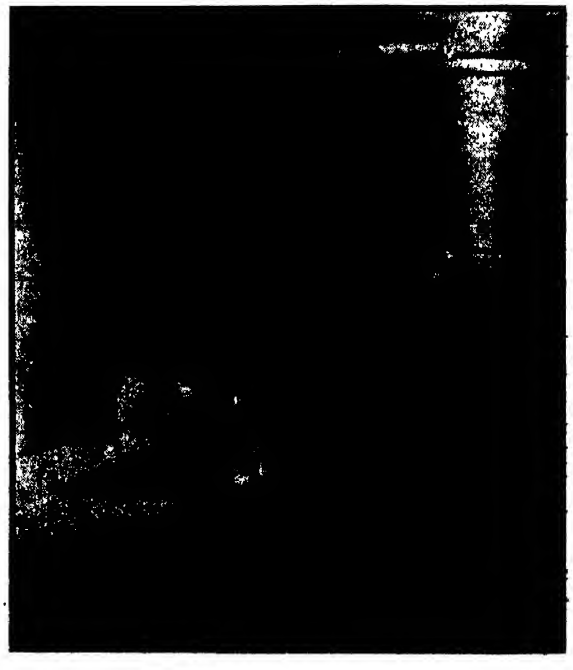


FIG. 2. A photograph showing the intermediate lens removed from the instrument.

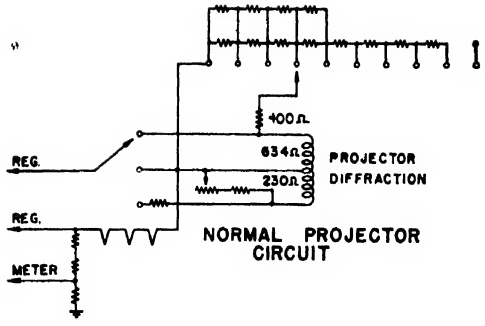


FIG. 3. A schematic diagram showing the circuit for the projection lens in the standard RCA Model EMU electron microscope.

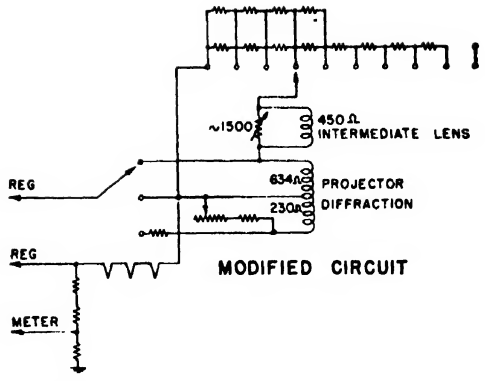


FIG. 4. A schematic diagram showing the circuit for the projection lens as modified for the insertion of the intermediate lens.

of the shunt resistor and the approximate magnifications achieved in the instrument using the standard short spacer in the projection lens polepieces.

The modified circuit used with the removable lens is shown in Fig. 4. The 400-ohm resistor in the arm circuit of the shunt system is replaced by the coil winding and another shunt resistor of rather higher value. The actual value of this second shunt resistor is set experimentally to attain the lowest magnification possible consistent with a full field of view and minimum distortion. In our model this value was 1500 ohms giving a minimum magnification of 990.

The performance of this system can be derived from the following simple theory. Figure 5 is a simplified diagram of the optical system including the intermediate lens. The following notation and values are used: $u_1 = -0.24$ cm, specimen to objective distance (fixed); v_1, image position for objective (cm); $u_2 = v_1 - d_1$, object position for intermediate lens (cm); $d_1 = 12.7$ cm, separation of objective and intermediate lens (fixed); $d_2 = v_2 = 11.4$ cm, image position for intermediate lens (fixed); f_1, focal length of objective (cm); f_2, focal length of intermediate lens (cm); f_3, focal length of projection lens (cm); m_3, magnification of projection lens having values given in Table I; m, magnification of objective and intermediate lenses combined; $M = mm_3$, total instrumental magnification.

At a given projection lens setting the magnification of the objective and intermediate lenses combined will determine the change in the total magnification as compared to the standard arrangement since the circuit change has virtually not affected the current flowing through the projection lens. Thus, to determine the total magnification for each step, it is necessary only to calculate the magnification of the objective-intermediate lens system and multiply it by the known value of the magnification of the projection lens at the same step. Thus

$$m = v_1 v_2 / u_1 u_2 = v_1 d_2 / u(v_1 - d_1) \qquad (1)$$

but

$$1/f_2 = 1/v_2 - 1/u_2 = 1/d_2 - 1/(v_1 - d_1) \qquad (2)$$

and

$$v_1 = \frac{f_2(d_1 + d_2) - d_1 d_2}{f_2 - d_2}. \qquad (3)$$

Substituting in expression (1) gives

$$m = \frac{d_1 + d_2}{u_1} \cdot \frac{d_1 d_2}{f_2 u_1} \qquad (4)$$

which on substituting numerical values gives

$$m = 100 - (600/f_2) \text{ (very nearly)}. \qquad (5)$$

From the expression (5) it is apparent that the focal length of the intermediate lens is the controlling factor as far as the magnification m is concerned. Furthermore, as the power $(1/f_2)$ of the intermediate lens increases, the over-all magnification first decreases to zero then increases to the limit of the lens. Since the power of the lens increases with increasing current, in the circuit described, it will have maximum power for Step 1 on the magnification control. Since the lowest magnification is desired at this step and the current is known, the design of the coil can be established. The maximum intermediate lens power that will be required is such that $m = 0$, that is, $f_2 = 6$ cm. Substituting numerical values in the equation for the focal length: $f_2 = (KVD_2)/(n_2 i_2)^2$ where K = instrumental constant (approximately 50); $V = 50{,}000$ volts; $D_2 = 1.27$ cm; n_2 = number of turns; i_2 = current in amp., shows that the coil should have 715 amp. turns. Since the current to be expected is 0.093 amp. approximately 8000 turns are required. By using No. 30 wire, such a coil will have a resistance of approximately 400 ohms. The maximum power consumption is 3.45 watts.

From the above data it is now possible to calculate the total magnification to be expected from the system. In these calculations, the projection lens polepiece with the short spacer is used. Furthermore, the calculations apply to the model of the intermediate lens built and installed with a 1500 ohm shunt resistor. The results of these calculations are collected in Table II. The calculated

TABLE II.

Tap no.	Int. lens current (ma)	m	M calc. at 45 kv K adj. at step 1	M meas. at 45 kv.
1	75.3	11.8	990	990
2	68.7	26.3	2630	2340
3	63.4	37.1	4350	3890
4	56.9	49.5	7130	6350
5	50.6	60.1	9500	9300
6	45.2	68.3	11750	11650
7	38.8	76.5	14500	14600
8	32.3	83.8	17200	17500
9	22.5	92.0	20400	21600
10	0	100.0	26000	26900

TABLE III.

Exposure no.	Magnification
1	978
2	990
3	995
4	1015

and measured values of M are plotted on Fig. 6. It can be seen that the agreement is quite good in spite of the numerous assumptions and approximations inherent in the calculations.

Magnifications below 900 are not satisfactory because of excessive barrel distortion (see Fig. 7). Actually, at a magnification of 900 a single opening of the 200-mesh screen commonly used as a specimen support just covers the 2×2 in. image field. Thus, no increase in field is achieved until more than one opening is imaged. This requires magnifications below 500 which are beyond the range of the present system. They can be achieved however, by using the intermediate lens in combination with other projection lens polepiece arrangements. In every case the maximum magnification given by a given projection lens polepiece remains unchanged and the minimum is 1/25 (approximately) of the maximum. The image field in each case is that of the specific projection polepiece arrangement.

A point which should be mentioned in connection with the use of this lens is that the variation in focus with change in magnification is no longer negligible. This is particularly true at the lower values of magnification. The dependence of f_1 on f_2 is given by the expression:

$$\frac{1}{f_1}=\frac{f_2-d_2}{f_2(d_1+d_2)-d_1d_2}\frac{1}{u_1}. \tag{6}$$

The substitution of numerical values shows that longer objective focal lengths are required for the lower magnifications. The increase for the entire range amounts to 7.1 percent most of which occurs in the lowest three steps. This corresponds to a reduction in objective current of approximately 3½ percent.

One other aspect of the optics of this system remains to be discussed, namely the distortion. Since for the range of values of f_2 considered the intermediate lens is always producing a real image of a virtual object, its spherical aberration results in a barrel distortion which becomes most apparent for those values of f_2 that cause the magnification m of the objective intermediate-lens system to approach zero. The projection lens, on the other hand, is always producing a real image of a real object and hence, as is well known, produces pincushion distortion. This is also more apparent at the lowest magnifications. Thus the distortion in the intermediate image is always opposite to that produced by the projector and tends to neutralize it.[4,5] It did not appear worth while to calculate the distortion since it was easier to measure it. Figure 7 shows the measured distortion at each magnification step. The distortion is expressed as a percentage increase in magnification, measured in the radial direction, at a point 2.5 cm from the center of the final image. It will be seen that for all practical purposes the system does not have excessive distortion. For comparison the distortion present in the final image produced by the standard projector lens with short spacer is plotted on the same figure.

IV. SOME PRACTICAL ASPECTS OF THE REMOVABLE INTERMEDIATE LENS

It is apparent from Fig. 6 that the relative rate of change of the total magnification M is very high for the lowest steps on the magnification control. For this reason, no attempt was made to design the intermediate lens for a specified minimum magnification. Instead, as already indicated, sufficient ampere-turns and resistance

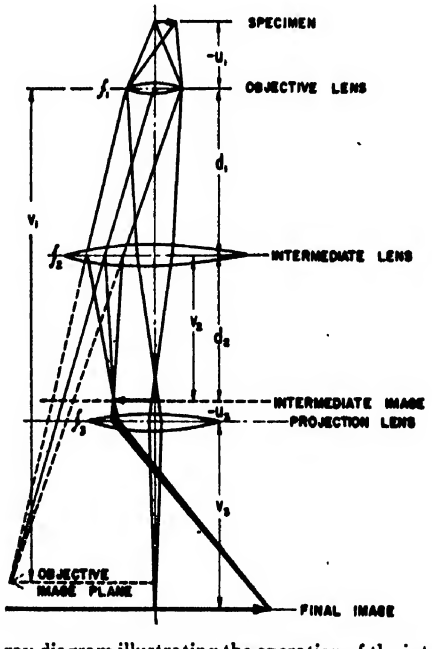


FIG. 5. A ray diagram illustrating the operation of the intermediate lens and the notation used in the text.

[4] J. Hillier, J. App. Phys. 17, 411–419 (1946).
[5] J. Hillier, U. S. Patent 2,418,349.

were built into the coil to permit adjustment by external shunts. The theory shows, and has been verified, that the magnifications obtained for settings above Step 2 were relatively insensitive to the value of this external shunt. In our work the shunt was adjusted so that one opening of 200-mesh screen was seen in its entirety on the final screen at Step 1 on the control. At this adjustment the high barrel distortion of the intermediate lens made itself apparent by a "folding" into the center of the image of screen openings *outside* the field being examined. This difficulty was eliminated by placing an aperture (0.150–0.200 in. diameter) in the intermediate lens.

The high rate of change of the magnification at the lower settings makes the system very sensitive to hysteresis effects in either the intermediate or projection

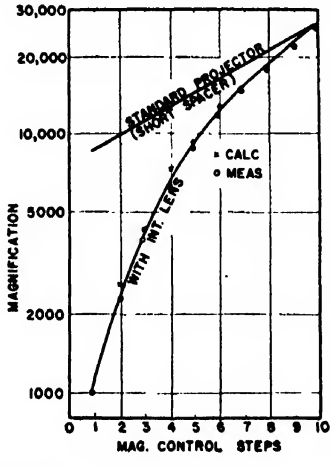


FIG. 6. A plot of the magnification against control position for the standard projector and for the standard projector in conjunction with the intermediate lens.

lenses. This was anticipated and particular care was taken in the choice of the iron used in the intermediate lens. When tested, however, it was found that the value of the lowest magnification was quite sensitive to the history of the lens. The extreme values encountered were 1110 and 990. The lower value was obtained by switching the projection lens circuit off and on several times. The higher value was obtained by leaving the circuit activated and turning the magnification control from Step 1 to Step 10 and back again several times. This manipulation showed that the effect was due to some retentivity in the iron of the projection lens since the two manipulations produce identical changes in the current flowing in the intermediate lens while they produce opposite changes in the current flowing in the projection lens.

In view of this situation, about which nothing could be done, some tests were made to check the reproducibility of the magnification at the lowest value. Four exposures were made of one field with the magnification at Step 1. Between the exposures the entire instrument

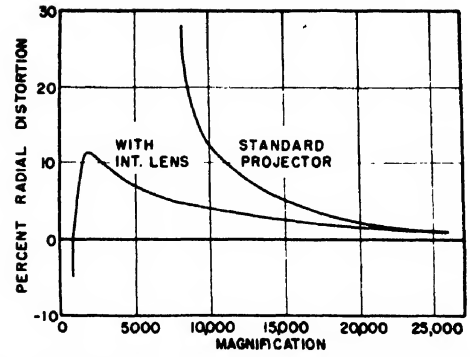


FIG. 7. A plot of percent radial distortion against magnification for the standard projector and for the standard projector in conjunction with the intermediate lens. The distortion is expressed as the percentage increase in magnification (measured in the radial direction) between the center and a point 2.5 cm from the center of the final image field.

was turned off and on and all the controls were changed in a random way. The image was then re-established and focused and the projection lens switch was turned off and on three or four times before making the exposure. The images obtained were then measured carefully. The results obtained are tabulated in Table III. It can be seen that the magnification had a mean value of 994 with a mean square deviation of 1.2 percent. Thus, while the system is sensitive to hysteresis effects, standardizing the magnetic history of the lens makes it possible to obtain very reproducible magnifications.

Since f_2 is proportional to the accelerating potential and the magnification given by Eq. (5) is a difference which approaches zero for the lowest step of the magnification control, it is apparent that the lowest magnifications are more sensitive to voltage changes than with the single projector. To check this, the percentage change in magnification which occurs when the accelerating potential was reduced from 50,000 volts to 45,000 volts was calculated for each step. The results are shown in Figs. 8 and 9; the former being a plot of the magnifications, the latter a plot of the percentage change in magnification. These calculations pointed out another important advantage of this system, namely that for Steps 4 and 5 (magnifications of 7140 and 9000) the *magnification of the system is essentially independent of*

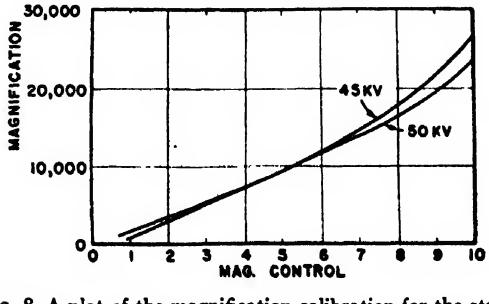


FIG. 8. A plot of the magnification calibration for the standard projector in conjunction with the intermediate lens for accelerating potentials of 45 kv and 50 kv.

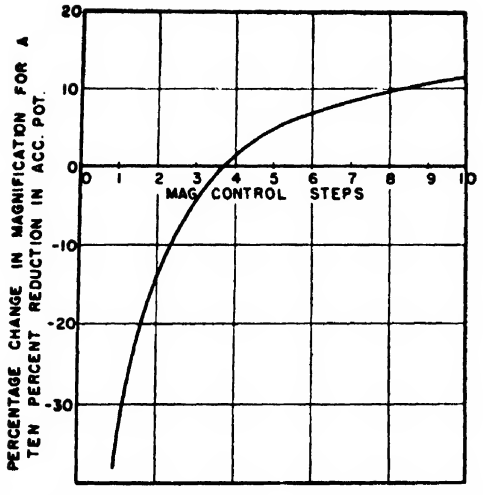


FIG. 9. A plot showing the percentage increase in magnification caused by a 10 percent reduction in accelerating potential.

the value of the accelerating potential. The potential change used in this example was 10 percent, which is many times greater than would be expected in normal use of the microscope, yet the change in magnification was only 1.5 percent for Step 5. Thus, if the specimen is placed accurately in the same position (within 0.001 in.) and the magnetic history of the intermediate and projection lens system is reproduced, the magnification calibration will remain accurate to within approximately 2 percent *for all time.* At all steps with the exception of Step 1, the dependence of the magnification on the accelerating potential (and hence on the state of the standard batteries) is the same or less than with a single projector.

The factors discussed in the preceding paragraph with regard to the dependence of magnification on the accelerating potential, also apply to changes in electron velocity caused by the absorption of energy in thick specimens. Thus, at Taps 4 and 5, chromatic change in magnification is minimized. This feature is particularly useful in the electron microscopy of sections of tissue. With uniformly thick 0.2-micron sections, the chromatic change in magnification has constituted a serious limitation of the field of view obtainable at the low magnifications used. To make the best use of the intermediate lens in the study of tissue sections, it is desirable to use a projection lens arrangement which gives a magnification of 700 to 1000 at Taps 4 and 5. The intermediate lens then becomes ideal for the examination of tissue sections.

A final point with regard to the practical use of the intermediate lens is the possibility of obtaining much lower magnifications for visual survey purposes by simply reducing the objective current. In experimental tests with thin specimens the defocusing could not be detected visually down to magnifications of 200 and was not serious to 100×. This does not apply to thick specimens such as tissue sections where the scattering greatly reduces the depth of field.

V. CONCLUSION AND SUMMARY

A removable intermediate lens has been described which extends the range of magnification in the direction of lower values without sacrificing the accessibility of the objective and projection lens pole pieces. Magnifications covering the range 900 to 24,000 are achieved without breaking the vacuum of the instrument. A complete field of view is obtained at all values and the distortion is in general less than with a conventional projector. The system provides one or two values of the magnification for which chromatic change in magnification is largely eliminated, thus providing the possibility of making an absolute and permanent magnification calibration.

While hysteresis effects in the existing projection lens cause the lowest values of the magnification to be sensitive to the magnetic history of the lens, it is shown that accurately reproducible magnifications can be obtained if the magnetic history is normalized.

The Wall-Quenching of Laminar Propane Flames as a Function of Pressure, Temperature, and Air-Fuel Ratio

RAYMOND FRIEDMAN AND W. C. JOHNSTON
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
(Received February 13, 1950)

The quenching distance between plane parallel plates has been measured for propane-air flames at eight pressures ranging from 0.0832 to 2.77 atmos., and air-propane ratios ranging from 11 to 24. Other tests have been made at atmospheric pressure in which (a) the temperature of the quenching plates has been varied from 80° to 715°F, (b) the temperatures both of the inlet gas and the quenching plates have been varied from 80° to 545°F, and (c) a series of six different solid surfaces was tested. The minimum quenching distance is found to occur at an air-propane ratio between 13.5 and 14.0, and is proportional to the minus 0.91 power of pressure. Lean mixtures are found to be slightly less pressure-sensitive. When both the inlet gas and plates are heated, the minimum quenching distance is found to be proportional to the minus 0.5 power of absolute temperature. The quenching effect appears to be independent of the nature of the surface.

INTRODUCTION

THE quenching effect of relatively cold solid surfaces upon pre-mixed laminar gas flames has been explored by several workers, including Holm,[1] Lewis, von Elbe *et al.*,[2–4] Friedman,[5] and Forsyth and Garside.[6] The aim of this investigation is to extend the data for the propane-air system, with particular emphasis on the pressure effect. Quantitative measurements of the effect of surface temperature and of inlet gas temperature are also made for the first time, and a search for catalytic surface effects is described.

The importance of the quenching effect in Bunsen flame flash-back has been brought out by Lewis and von Elbe. The possibility exists that quenching plays a role in blow-off from a flame-holder, although this has not been conclusively demonstrated. Williams, Hottel, and Scurlock[7] found that heating of the flame-holder extends the stability limits considerably, which suggests that quenching by the cold anchor is of some importance. Particularly at low pressures, where the quenching distance is large, one might expect that incomplete combustion near cold burner walls is associated with quenching. Further, it appears reasonable that rich and lean inflammability limits and the low pressure limit of inflammability are governed by quenching effects. To illustrate this, the minimum pressure at which gasoline vapor-air may be ignited in a 2-in. tube is reported by Jones and Spolan[8] as 0.046 atmos. The extrapolated results of this investigation show the minimum quenching distance of propane-air at 0.046 atmos. to be 1.2 in. for plane-parallel walls, and this would be about 35 percent greater in a tube (or 1.6 in.), according to Fig. 6 of reference 3. This 1.6-in. quenching diameter for propane may be compared with the 2-in. tube in which gasoline ceases to propagate flame at this pressure, and it is seen that quenching must be important in burning near the low pressure limit. At atmospheric pressure also, the rich and lean limits as determined by upward propagation in tubes appear to depend upon quenching; data obtained by White[9] for five fuels (H_2, CO, CH_4, C_2H_4, C_2H_2) and air in 1-, 2-, and 3-in. tubes show a progressive broadening of both limits in each case with increasing tube size. The broadening of these limits may be thought of as extensions of the limbs of the familiar U-shaped curve of quenching distance against air-fuel ratio.

This investigation was undertaken, therefore, because of the general importance of the quenching phenomenon in combustion. Measurements were made by measuring the minimum spacing between plane-parallel plates through which flash-back can occur with a given fuel-air ratio, pressure, ambient temperature, and wall temperature. Propane was chosen as the fuel because of availability in pure form, ease in metering, and the

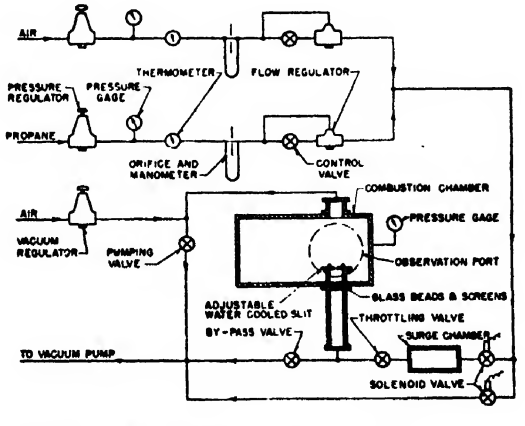


FIG. 1. Apparatus for measuring quenching distances at various pressures.

[1] J. M. Holm, Phil. Mag. 14, 18 (1932); 15, 329 (1933).
[2] Blanc, Guest, von Elbe, and Lewis, J. Chem. Phys. 15, 798 (1947).
[3] Harris, Grumer, von Elbe, and Lewis, *Third Symposium on Combustion and Flame and Explosion Phenomena* (Williams and Wilkins, Baltimore, 1949), p. 80.
[4] Blanc, Guest, von Elbe, and Lewis, see reference 3, p. 363.
[5] R. Friedman, see reference 3, p. 110.
[6] J. S. Forsyth and J. E. Garside, see reference 3, p. 99.
[7] Williams, Hottel, and Scurlock, see reference 3, p. 21.
[8] G. W. Jones and I. Spolan, U. S. Bureau of Mines R. I. 3966 (1946).
[9] A. G. White, J. Chem. Soc. 125, 2387 (1924).

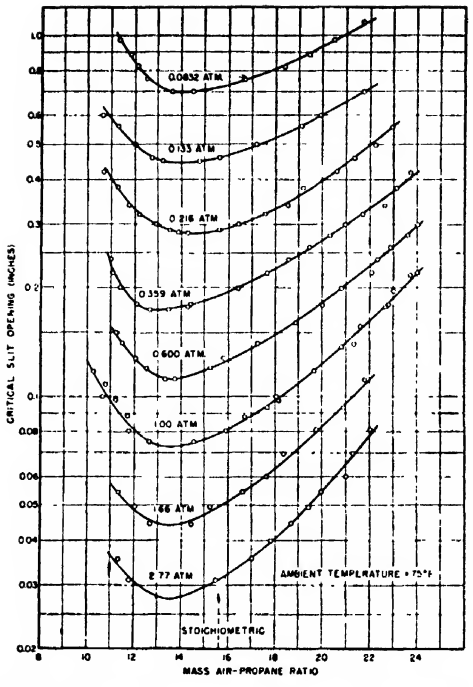


FIG. 2. The quenching of propane-air flames at various pressures

similarity between propane and higher saturated hydrocarbons with regard to burning velocity,[10] minimum ignition energy,[3] flame temperature, and low temperature oxidation characteristics.[11]

EXPERIMENTAL TECHNIQUE

The procedure for determining the quenching distance consists in producing a flame stabilized on a rectangular port whose length-to-width ratio is always greater than 3.6. When the flow to the burner is quickly reduced to zero the flame will either be quenched by the plates or will flash back through the slit, depending on the air-fuel ratio of the mixture and the ambient conditions. The critical mixture is determined by testing a series of mixtures until a composition is found that separates a region of flash-back from a region of quenching. In general, there is a rich as well as a lean critical mixture corresponding to any particular set of conditions.

Variable-Pressure Flow Apparatus

The apparatus for measuring the quenching distance at various pressures is shown schematically in Fig. 1. A mixture of air and propane of controlled composition and pressure is supplied to the combustion chamber by a flow system. A thermocouple connected to a microammeter indicates the occurrence of flash-back, which can also be observed through a Pyrex window. The flow is reduced to zero by the action of the solenoid valves, surge chamber, and throttling valve. It was found that the sudden movement of the plunger in the solenoid valve was sufficient to disturb the flame, so that a surge chamber and throttling valve were necessary to smooth the flow during this period. The throttling valve was set to reduce the flow to zero in about 0.1 to 0.3 sec., although this setting was not found to be critical, and the data were reproducible over a fairly wide range of valve settings.

The vertical approach section to the burner is four inches in diameter and contains glass beads held in position by 60-mesh copper screens. The jaws of the slit are $\frac{3}{16}$-in. copper plates with water-cooling tubes attached. Slit widths are adjusted by thickness gauges.

The propane* used for these tests is reported to be 99.0 percent pure. The air is supplied at 80 p.s.i. g and contains 0.33 percent by weight water vapor. The effect of this concentration of water vapor upon the data appears to be negligible since the atmospheric curve of Fig. 2 agrees with previous work done with dry air.

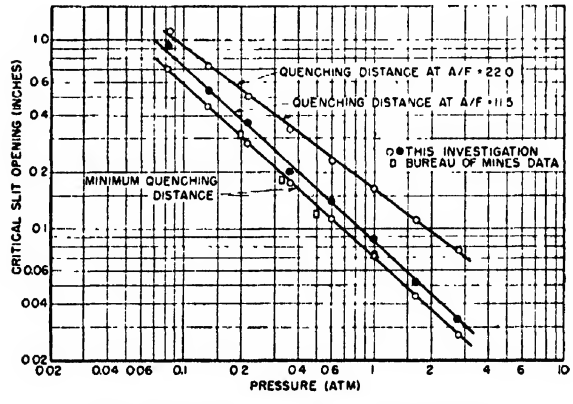


FIG. 3. The dependence of propane-air flame quenching on pressure.

The flow meters were calibrated against a Sargent wet-test meter.

Variable-Temperature Flow Apparatus

The variable-temperature apparatus consists of identical air and propane flow systems. In each system the gas from a tank is reduced by a regulator and supplied to a critical-flow orifice.[12] The temperature and pressure at the orifice determine the density and hence the mass flow of gas. The pre-mixed gases are supplied to the burner through a hand-operated valve which is used for reducing the flow when testing. The burner consists of parallel copper plates with independent means for electrical heating of the plates and the gas. Tests showed that the results are independent of the plate thickness over the range from $\frac{1}{16}$ to $\frac{3}{16}$ in., as long as the inlet gases and the plates are at the same tempera-

[10] T. W. Reynolds and M. Gerstein, see reference 3, p. 190.
[11] K. Spence and D. T. A. Townend, see reference 3, p. 404.

* Matheson Company, East Rutherford, New Jersey.
[12] J. W. Andersen and R. Friedman, Rev. Sci. Inst. **20**, 61 (1949).

ture. Thermocouples are embedded in the plates for measuring the temperature of the quenching surface. Results indicate that the temperature gradient in the copper is negligible, being about 1°F per $\frac{1}{8}$ in. near the quenching surface when the flame is burning and the plates are water-cooled.

Factors Limiting Extent of Data

It is of interest to discuss the factors limiting the range of testing at the highest and lowest pressures and at the richest and leanest compositions. On the rich side, when air-propane ratios were below about 11, moving ridges appeared in the flame surface, often in a herringbone pattern. On further enrichment of the mixture, the flame would resolve itself into several individual cellular flames, each convex toward the unburned gas. Similar examples of polyhedral and cellular flames have been described in the literature.[13,14] When attempts were made to quench these flames, erratic results were obtained. Since these rich mixtures are characterized by very low burning velocities, it appears that sufficient time is available for surrounding gases to mix with the combustible gas beneath the flame surface.

On the lean side, it became progressively more difficult to ignite and stabilize the flames, the tendency for blow-off being very pronounced. At the lowest pressure the primary difficulty was ignition; it was found necessary to ignite each mixture at about 0.13 atmos. and then gradually reduce the pressure to 0.0832 atmos. for testing. This pressure represents substantially the lowest that can be obtained with the available pumping equipment. At the highest pressure, 2.77 atmos., and slit openings of 0.030 in. and less, testing was difficult because of excessive moisture condensation from the flame upon the plates.

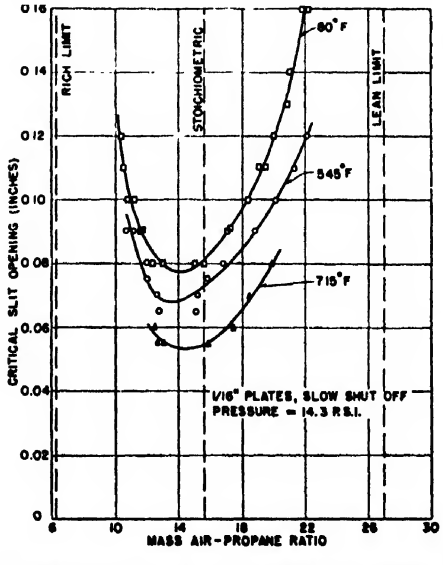


FIG. 4. Quenching of propane-air flames with various plate temperatures (gas initially 80°F).

[13] F. A. Smith and S. F. Pickering, Bur. Stand. J. Research 3, 65 (1929).
[14] G. H. Markstein, J. Chem. Phys. 17, 428 (1949).

TABLE I. The quenching of propane-air flames at various pressures, at 75°F.

Pressure (atmos.)	Slit opening (in.)	Critical mass air-propane ratio Lean	Rich	Pressure (atmos.)	Slit opening (in.)	Critical mass air-propane ratio Lean	Rich
0.0832	0.700	14.41	13.52	1.00	0.075	14.52	12.61
0.0832	0.760	16.64	12.48	1.00	0.0804	15.89	11.77
0.0832	0.822	18.32	12.06	1.00	0.088	16.70	11.71
0.0832	0.890	19.37	11.78	1.00	0.093	17.69	—
0.0832	0.980	20.40	11.29	1.00	0.098	18.15	11.20
0.0832	1.110	21.65	—	1.00	0.100	18.05	10.65
0.133	0.450	14.71	13.13	1.00	0.108	—	10.78
0.133	0.460	15.57	12.70	1.00	0.1175	19.63	10.25
0.133	0.500	17.13	11.98	1.00	0.137	20.80	—
0.133	0.560	19.05	11.26	1.00	0.140	21.30	—
0.133	0.600	19.87	10.58	1.00	0.157	21.60	—
0.133	0.700	21.70	—	1.00	0.177	22.70	—
0.216	0.285	14.22	13.82	1.00	0.180	22.80	—
0.216	0.290	15.54	13.47	1.00	0.197	23.00	—
0.216	0.300	16.38	12.89	1.00	0.2165	23.70	—
0.216	0.320	17.51	12.17	1.00	0.220	24.00	—
0.216	0.340	18.50	11.73	1.66	0.0455	14.41	12.68
0.216	0.380	19.15	11.25	1.66	0.0495	15.22	12.05
0.216	0.420	20.55	10.65	1.66	0.0545	16.63	11.32
0.216	0.460	21.35	—	1.66	0.060	17.63	—
0.216	0.500	22.20	—	1.66	0.0695	18.35	—
0.216	0.560	22.90	—	1.66	0.081	19.74	—
0.359	0.174	13.42	12.63	1.66	0.093	20.90	—
0.359	0.177	14.24	—	1.66	0.111	21.80	—
0.359	0.180	14.39	12.08	2.77	0.031	15.47	11.80
0.359	0.200	16.40	11.39	2.77	0.0355	17.01	11.32
0.359	0.220	17.60	11.04	2.77	0.040	17.81	—
0.359	0.240	18.51	10.96	2.77	0.0455	18.70	—
0.359	0.260	19.42	—	2.77	0.0495	19.42	—
0.359	0.280	20.30	—	2.77	0.0545	19.95	—
0.359	0.300	20.95	—	2.77	0.060	21.00	—
0.359	0.320	21.70	—	2.77	0.0695	21.30	—
0.359	0.340	22.60	—	2.77	0.081	22.04	—
0.359	0.380	23.10	—				
0.359	0.420	23.70	—				
0.600	0.112	13.70	13.30				
0.600	0.120	15.21	12.50				
0.600	0.128	15.79	12.02				
0.600	0.140	17.22	11.47				
0.600	0.150	—	11.23				
0.600	0.160	18.88	—				
0.600	0.180	20.00	—				
0.600	0.200	20.80	—				
0.600	0.220	22.05	—				
0.600	0.240	22.30	—				
0.600	0.260	22.85	—				
0.600	0.280	23.60	—				
0.600	0.300	24.00	—				

In every case, data were taken over as wide a range as was feasible with the present apparatus.

RESULTS

Precision and Accuracy

The data showing the effect of pressure upon quenching appear in Fig. 2. There are 118 points, each representing a group of bracketing tests; 78 percent of these points are within one percent of the curves, and 97 percent are within three percent of the curves. No points have been discarded. The points were not obtained in any regular sequence. Thus, the precision, or

TABLE II. The quenching of propane-air flames with various inlet temperatures, at 14.3 p.s.i. pressure (gas and plates at same temperature).

Temperature (°F)	Slit opening (in.)	Critical mass air-propane ratio Lean	Rich
80	0.080	14.95, 15.62	12.98, 12.32
80	0.090	17.27, 17.06	11.52, 11.71
80	0.100	18.35, 18.47	11.11, 10.68
80	0.110	19.15, 19.30	10.47
80	0.120	20.00	10.26
80	0.130	20.85	—
80	0.140	20.99	—
80	0.160	22.15, 21.75	—
260	0.070	14.84	13.14
260	0.075	16.08	12.20
260	0.085	18.08	11.38
260	0.090	18.85	11.10
260	0.100	19.81	—
260	0.120	21.57	—
410	0.060	14.84	13.35
410	0.065	16.08	12.53
410	0.075	18.37	11.46
410	0.085	20.13	10.80
410	0.090	20.64	10.52
410	0.100	21.75	—
410	0.110	22.83	—
545	0.057	15.23	13.45
545	0.058	16.60	12.94
545	0.065	18.56	11.54
545	0.075	20.13	11.01
545	0.080	21.24	10.66
545	0.085	21.71	10.33

reproducibility, of the test procedure is seen to be satisfactory.

An indication of accuracy was obtained by determining the quenching curve at 1 atmos. and 75°F both with the variable-pressure and the variable-temperature apparatus, these two flow systems having been independently calibrated. Also the burner geometry and method of flow reduction to produce flash-back were somewhat different in the two cases. Upon comparing the two curves at a given slit opening, the air-propane ratios never differed by more than two percent. The minimum slit openings differed by 5.5 percent.

The Bureau of Mines investigators[4] report data for minimum spacing of flanged electrodes for ignition in a static system with a spark of minimum energy. Their data for propane-air between 0.2 and 1 atmos. agree very well with the results of this investigation, as shown in Fig. 3.

Effect of Pressure and Composition

In Table I and Fig. 2, data are presented showing critical slit openings for quenching of propane-air flames from 0.0832 to 2.77 atmos., at 75°F ambient temperature. At each pressure, a U-shaped curve is obtained with the minimum at a mass air-propane ratio near 14 at the lower and 13.5 at the higher pressures. The stoichiometric air-propane ratio is 15.65. The minima are plotted against pressure in Fig. 3, a log-log plot, and a straight line is obtained, from the slope of which it follows that the minimum quenching distance is proportional to the minus 0.91 power of the pressure. It may be noticed in Fig. 2 that the pressure effect upon quenching is slightly smaller for the leaner mixtures. Upon plotting in Fig. 3 the quenching distance for an air-propane ratio of 22.0 against pressure, a straight line is again obtained, of reduced slope. The quenching distance is proportional to the minus 0.76 power in this case. For a mixture on the rich side of the minimum (air-propane ratio of 11.5), the minus 0.91 power is still valid.

The Bureau of Mines spark-quenching data for propane-air are seen to be consistent with these results both in magnitude and in pressure variation. Further, the Bureau of Mines spark-quenching data for methane-air between 0.1 and 1 atmos. reveal a minus 0.88 power pressure dependence when the minima are plotted in this fashion.

Effect of Temperature and Composition

In the first series of temperature tests at atmospheric pressure, the quenching plates were electrically heated to various temperatures while the inlet gas was maintained at 80°F. Results are shown in Fig. 4. It is seen that the quenching effect is diminished by increasing the plate temperature, as would be expected. In tests of this type, a certain unavoidable heating of the gas occurs as it flows through the slit. Rough measurements showed that the gas was heated to about 180°F when the plates were at 540°F. Because of this effect, the data of Fig. 4 are only semiquantitative.

Other tests were made at atmospheric pressure in which both the gas and the plates were heated to the same temperature, with results shown in Fig. 5 and Table II. The quenching effect is seen to become regularly less as the temperature is increased. For rich and stoichiometric flames, the quenching distance varies with about the

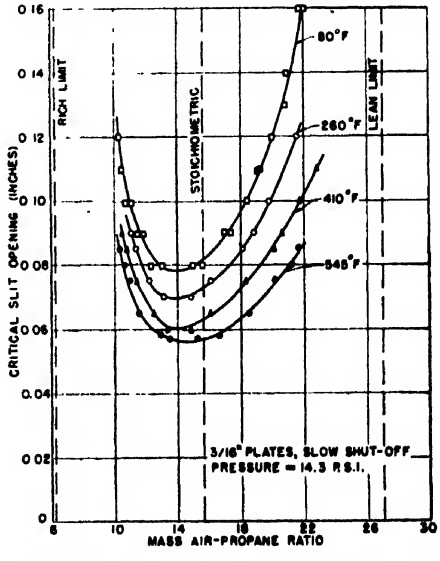


FIG. 5. Quenching of propane-air flames with various inlet temperatures (gas and plates at same temperature).

minus 0.5 power of absolute temperature, while for lean flames the negative exponent increases, being 0.85 at an air-propane ratio of 21.0.

Effect of Surface Coatings

A series of surface coatings on copper was investigated, with the gas and $\frac{3}{16}$-in. plates at 410°F. No measurable differences in quenching behavior were observed for either lean or rich mixtures, for any of the following types of surfaces: nickel, platinum, bright gold, roughened gold, potassium chloride, and water glass. This result indicates that neither the catalytic power of the surface nor its absorptivity for radiation significantly affects the quenching phenomenon.

Electromagnetic Field of the Conical Horn*

MARVIN G. SCHORR** AND FRED J. BECK, JR.
Yale University, New Haven, Connecticut
(Received January 20, 1950)

Maxwell's equations are solved for a perfectly conducting conical wave guide and the propagation coefficients of this guide are discussed. The field at the mouth of a finite conical horn is estimated and the radiation therefrom is calculated in integral form. These integrals are solved in series form for horns of small flare angle and moderate length and the results compared with experiment.

INTRODUCTION

THE use of electromagnetic horns for both matching purposes and enhanced radiation directivity in high frequency antennae was suggested by Southworth[1] and Barrow[2] in 1936. Since that time many observers have verified these properties and several attempts have been made to attain a satisfactory agreement between theory and experiment. The early workers calculated radiation phenomena from the scalar Kirchhoff-Huygens principle, but it was soon recognized that this formalism was inadequate, the major defect being in the assumption of continuity of the fields at the aperture from which the radiation ensues. Two rigorously correct methods are available for computing radiation from horn structures. They are the vector diffraction theory of Stratton and Chu,[3] and the assumed field technique recently restated by Schelkunoff.[4] A recent paper by Woonton, Hay, and Vogan[5] gives a good discussion of some of these matters.

The great majority of the existent theoretical work has been done for propagation in and radiation from uniform wave guide and sectoral horns. Buchholz[6] gave a rather complete analysis of electromagnetic fields in conical wave guide and also commented on a method of exciting such a guide. However he treated only the special case of azimuthally symmetric guided waves. Excellent experimental work is available in the literature on both conical[7] and pyramidal horns[8] and it is the purpose of this paper to show that the rigorous assumed-field technique, when applied to the conical horn, yields good agreement with experiment over a range of cone angles for which the radiation integrals may be satisfactorily evaluated.

FIELDS IN CONICAL WAVE GUIDE

Let us first consider the permissible electromagnetic fields inside a perfectly conducting conical wave guide of semi-infinite extent, filled with a dielectric of permeability μ and permitivity ϵ, ϵ being complex if the dielectric is lossy. The choice of a spherical (r, θ, ϕ) coordinate system is clearly indicated, and if the apex of the cone coincides with the origin of coordinates and the polar axis is the longitudinal axis of symmetry of the cone, then conical surfaces will coincide with surfaces of constant colatitude and the flare angle of the cone will be measured by the colatitude of its surface. The flare angle, so defined and henceforth used, is then one-half the total angle subtended by the cone. The cone is assumed excited near the apex in such a manner that the fields propagate positively radially outward; also the azimuthal reference axis is chosen so that a single trigonometric function describes the azimuthal variations of the fields. The permissible fields are well known and are given by:

Transverse-Electric Waves

$$
\begin{aligned}
E_r &= 0 \\
E_\theta &= Bj\omega\mu \cos m\phi h_n^{(2)}(kr) P_n^m(x) \\
E_\phi &= -Bj\omega\mu \sin\theta \sin m\phi h_n^{(2)}(kr) P_n^{m\prime}(x) \\
H_r &= B[n(n+1)/r] \sin m\phi h_n^{(2)}(kr) P_n^m(x) \\
H_\theta &= -B \sin\theta \sin m\phi [(1/r) h_n^{(2)}(kr) \\
&\qquad + k h_n^{(2)\prime}(kr)] P_n^{m\prime}(x) \\
H_\phi &= B[m/\sin\theta] \cos m\phi [(1/r) h_n^{(2)}(kr) \\
&\qquad + k h_n^{(2)\prime}(kr)] P_n^m(x);
\end{aligned}
\tag{1}
$$

* This paper is an abstract of a dissertation presented by the first named author (Marvin G. Schorr) for the degree of Doctor of Philosophy in Yale University, May, 1949.

** Now at Tracerlab, Inc., Boston, Massachusetts.

[1] G. C. Southworth, Bell Sys. Tech. J. 15, 284 (1936).
[2] W. L. Barrow, Proc. I.R.E. 24, 1298 (1936).
[3] J. A. Stratton and L. J. Chu, Phys. Rev. 56, 92 (1939).
[4] S. A. Schelkunoff, Phys. Rev. 56, 308 (1939).
[5] Woonton, Hay, and Vogan, J. App. Phys. 20, 71 (1949).
[6] H. Buchholz, Ann. d. Physik 37, 173 (1940).
[7] G. C. Southworth and A. P. King, Proc. I.R.E. 27, 95 (1939).
[8] D. R. Rhodes, Proc. I.R.E. 36, 1101 (1948).

Transverse-Magnetic Waves

$$
\begin{aligned}
E_r &= B[n(n+1)/r]\sin m\phi h_n^{(2)}(kr)P_n^m(x)\\
E_\theta &= -B\sin\theta\sin m\phi[(1/r)h_n^{(2)}(kr)\\
&\qquad +kh_n^{(2)\prime}(kr)]P_n^{m\prime}(x)\\
E_\phi &= B[m/\sin\theta]\cos m\phi[(1/r)h_n^{(2)}(kr)\\
&\qquad +kh_n^{(2)\prime}(kr)]P_n^m(x)\\
H_r &= 0\\
H_\theta &= B(j\omega\epsilon m/\sin\theta)\cos m\phi h_n^{(2)}(kr)P_n^m(x)\\
H_\phi &= Bj\omega\epsilon\sin\theta\sin m\phi h_n^{(2)}(kr)P_n^{m\prime}(x),
\end{aligned}
\tag{2}
$$

where the primed function always denotes the derivative with respect to the argument, and the wave number $k=2\pi/\lambda$, λ being the free space wave-length and $\omega^2=\mu\epsilon$. The $h_n^{(2)}(kr)$ and $P_n^m(x)$ are the spherical Hankel function of the second kind and associated Legendre function of the first kind respectively, and $x=\cos\theta$. The spherical Hankel function of the first kind does not appear since it represents a wave traveling in the negative radial direction, contrary to hypothesis; the associated Legendre function of the second kind also does not appear since it has a singularity on the polar axis which is contained in the region considered. The spherical Hankel function of the second kind has a singularity at the origin but this is of no consequence since, as will be seen later, the neighborhood of the origin is very highly attenuating so that no energy ever reaches it. The single-valuedness of the solutions in azimuth requires that m be an integer which we may take as positive. The colatitude boundary condition, namely $E_r=E_\phi=H_\theta=0$ at $\theta=\theta_0$, requires for TE waves,

$$[-\sin\theta P_n^{m\prime}(\cos\theta)]_{\theta=\theta_0}=0, \tag{3}$$

and for TM waves,

$$P_n^m(\cos\theta_0)=P_n^m(x_0)=0. \tag{4}$$

Hence n must be a root of the multibranched transcendental Eqs. (3) or (4) considered as functions of the order n of the associated Legendre function. In general n will be non-integral and a function of the flare angle θ_0 of the cone; we may consider only the positive roots. For small integral values of n, $0<n\leqq 10$, excellent tables of associated Legendre functions are available for determining the zeros of Eqs. (3) and (4)[9] and reasonably accurate interpolation to non-integral orders is possible. For large orders the zeros may be determined from the asymptotic expansions of these functions.

The solutions of Eqs. (1) and (2) then form a doubly infinite set corresponding to all positive integral m and all positive n giving zeros of Eqs (3) and (4). Those modes actually present in the guide will of course be determined by the source of excitation.

It is convenient in discussing conical wave guide to adopt the cylindrical wave guide terminology since if we allow a given cross section of the cone to remain constant and let the flare angle approach zero, then in the limit the cone goes over into a hollow cylinder. The permissible modes in conical guide are hence designated TE_{mp} or TM_{mp} having m field variations in one azimuthal circuit and p field variations from the polar axis to the cone wall. The designation of propagation constants for conical guide does not, however, follow in the usual sense. In conical wave guide the fields are not periodic functions of the propagation coordinate r and propagation constant loses much of its intuitive meaning. However, in lieu of a better definition, a complex propagation coefficient is defined for conical wave guide, in the manner used by Barrow and Chu for sectoral horns,[10] by

$$\gamma(Q,r)=-1/Q(\partial Q/\partial r)=-(\partial/\partial r)(\log Q), \tag{5}$$

where Q is any field component. In general γ will be a function of r and will be different for various field components. This definition, it may be observed, reduces directly to the usual definition of the propagation constant if Q is simple harmonic in r, and γ is then a true constant. For regions in which γ is not too rapid a function of radius useful information is still obtained.

In an entirely similar manner we may define a pseudo guide wave-length $\lambda_g(Q,r)=2\pi/Im\gamma$, and a pseudo phase velocity $v_p(Q,r)=\omega/Im\gamma$. From Eq. (1) for TE waves and the asymptotic expansions of the spherical Hankel function for large and small argument

$$
h_n^{(2)}(kr)=\begin{cases}\left(\frac{\pi}{4}\right)^{\frac{1}{2}}\dfrac{1}{(n+\frac{1}{2})!}\left(\dfrac{kr}{2}\right)^n\\ \qquad +j\left(\dfrac{\pi}{4}\right)^{\frac{1}{2}}(n-\frac{1}{2})!\left(\dfrac{2}{kr}\right)^{n+1}, & kr\ll 1\\ \dfrac{1}{kr}\exp\left[-j\left(kr-\dfrac{n+1}{2}\pi\right)\right], & kr\gg 1\end{cases}
\tag{6}
$$

where $(n+\frac{1}{2})!$ has been written for the gamma function $\Gamma(n+\frac{3}{2})$, we may obtain the parameters of Table I which gives the propagation coefficient, the pseudo guide wave-length, and the pseudo phase velocity for the E_θ and H_ϕ components. Following Barrow and Chu we designate $kr\gg 1$ as a transmission region and that for which $kr\ll 1$ an attenuation region. In the transmission region the waves behave asymptotically as spherical waves in a free medium while in the attenuation region the behavior is similar to that in a uniform cross section guide near the cut-off wave-length.

These results can be expected to have physical validity only if the propagation coefficient γ is not too severe a function of r, or in other words, if the radial space variation is almost periodic. In the region near the apex where this condition is not true two anomalies arise. First, the phase of $\gamma(H_\phi)$ in the attenuation region is negative, which in uniform guide indicates a wave traveling in the negative radial direction, contrary to

[9] Math. Tables Project, Columbia University Press, New York (1945).

[10] W. L. Barrow and L. J. Chu, Proc. I.R.E. 27, 51 (1939).

hypothesis. Second, the group velocity v_g, if computed from the usual relation $v_g = d\omega/d(I_m\gamma)$ turns out to be of the same form as the phase velocity and the product of group and phase velocities is not c^2, as is required for a dispersionless medium. These anomalies can be ascribed to the severe aperiodicity of the Hankel functions in the region near the apex. It should be emphasized that the physical significance of the propagation coefficients here developed should not be exaggerated; however the implications of the results are of value in understanding the phenomena occurring.

Additional light on the properties of conical wave guide may be obtained from the behavior of the Poynting vector and power flow in the guide. It is not difficult to show that the angular components of the Poynting vector S are always pure imaginaries whereas the radial component is in general complex. In the attenuating region the radial component is largely imaginary along with the angular components. These latter decrease as $1/r^3$ at large distances and soon reach a negligible value compared with the radial component which decreases only as $1/r^2$.

From Eqs. (1) and (2), and the asymptotic expansions of the Hankel functions, Eq. (6), it is readily observed that in the transmission region all transverse field components decrease with reciprocal radius while the radial components decrease as reciprocal radius squared. As the wave progresses down the guide it therefore becomes more nearly a TEM wave matched to free medium, as may be seen from the wave impedance. For TE and TM waves, respectively, the wave impedance is given asymptotically by

$$kr \gg 1, \quad \begin{cases} Z_E = (\mu/\epsilon)^{\frac{1}{2}} \\ Z_H = (\mu/\epsilon)^{\frac{1}{2}} \end{cases}$$

$$kr \ll 1, \quad \begin{cases} Z_E = (\mu/\epsilon)^{\frac{1}{2}} \left[\dfrac{\pi(2n+1)}{n(n-\frac{1}{2})!(n+\frac{1}{2})!} \times \left(\dfrac{kr}{2}\right)^{2n+2} + j\dfrac{4}{nkr} \right] \\ Z_H = (\mu/\epsilon)^{\frac{1}{2}} \left[\dfrac{\pi(2n+1)}{(n-\frac{1}{2})!(n+\frac{1}{2})!} \times \left(\dfrac{kr}{2}\right)^{2n} - j\dfrac{n}{kr} \right]. \end{cases} \quad (8)$$

In regions near the apex the wave impedances are largely reactive, inductive for TE and capacitive for TM waves. As the wave moves into the transmission region the reactive part decreases, the resistive part increases and approaches asymptotically the characteristic impedance $(\mu/\epsilon)^{\frac{1}{2}}$ of the free medium.

RADIATION FROM A CONICAL HORN

Let us consider a cylindrical wave guide feeding a concentric conical wave guide. We will assume the cylindrical guide is excited in its lowest mode, TE_{11}, as is usually the case. At the junction of the uniform guide with the cone there is a discontinuity in the slope of the conducting boundary and higher order modes are required to satisfy the boundary conditions, the fundamental mode being partially reflected and partially transmitted. In the cylindrical guide the higher order modes will be rapidly attenuated but the behavior in the conical guide depends on whether the waves are introduced in the attenuating or transmitting regions of the cone. Certainly the cone will be transmitting for the fundamental mode on purely physical grounds since the diameter of the cone cross section is always greater than that of the uniform guide. The higher order modes, however, may be in an attenuating region and we have seen that such modes have an attenuation coefficient linearly proportional to the mode order, n, of the wave. The distance that a given order mode travels in an attenuating region may be calculated in a strictly qualitative manner, if it is assumed that in conical guide any order wave is in an attenuating region until the cross-sectional diameter of the cone approximates the diameter of a cylindrical guide just permitting propagation of that order wave in the uniform guide. If the diameter of the uniform guide operating in its lowest mode is D_0, and if the diameter corresponding to the nth mode is D_n, then the distance of travel of the nth mode wave in an attenuating region of the conical guide of flare angle θ_0 may be taken as

$$d = (D_n - D_0)/(2\sin\theta_0).$$

Small flare angles result in long journeys through attenuating regions. Also small flare angles result in smaller discontinuities at the cylinder-cone junction and, hence, require less higher mode energy to satisfy the boundary conditions. These factors both markedly discriminate against higher order modes.

In the conical guide, then, there will be present TE and/or TM waves composed principally of a fundamental and contaminated with higher order modes whose magnitudes depend on the flare angle and length of the guide. As the wave propagates in the positive radial direction, Eqs. (1), (2), and (6) show that, asymptotically, the radial components decrease only as reciprocal radius. Hence, some distance down the conical guide the wave will be almost transverse electromagnetic with only a small radial component. As this occurs, the wave impedance approaches that of a wave in free medium. Cutting off the conical guide at this distance should not materially alter the fields anywhere inside the cone since it is essentially matched into free space; very little energy will be reflected from the mouth. The small radial component still present is just sufficient to close the loops of the field lines of the TEM wave in free medium. The open ended cone thus becomes a source of radiation.

The calculation of the radiation characteristics of such a source involves two approximations at the outset:

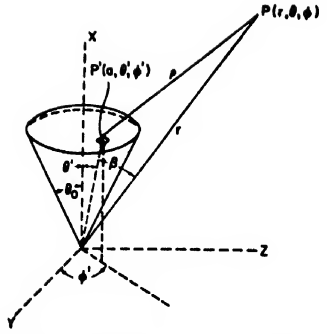


FIG. 1. Coordinate system for radiation calculations.

(1) the field at the mouth of the horn is assumed to be that which would exist at that cross section if the horn extended to infinity; and (2) there is assumed to be no field leakage around the rim of the mouth to the exterior surface of the horn. These matters have been discussed at length by others and will here be assumed valid provided the cone flare angle is not too great and its length is not too short.

The assumed field technique of Love,[11] MacDonald,[12] and later Schelkunoff[4] leads to the following expressions for the electric and magnetic intensities.

$$\begin{aligned} E&=k^2A+\nabla\nabla\cdot A-j\omega\mu\nabla\times F,\\ H&=k^2F+\nabla\nabla\cdot F+j\omega\epsilon\nabla\times A, \end{aligned} \tag{9}$$

where the vector potentials A and F are given by

$$\begin{aligned} F&=-\frac{1}{j\omega\mu}\int_{S'}\frac{[n\times E']ds'}{4\pi\rho},\\ A&=\frac{1}{j\omega\epsilon}\int_{S'}\frac{[n\times H']ds'}{4\pi\rho}. \end{aligned} \tag{10}$$

and E and H are the fields at any point in space distant ρ from a surface element ds' on which a field E' and H' exist, n being the unit vector normal to the surface and brackets indicating retardation in the Lorentz sense. The TE field at the mouth of the horn, S', is given by Eq. (1), and Fig. 1 shows the coordinate system used. Note that primed coordinates refer to the cone mouth while unprimed coordinates refer to the field point $P(r, \theta, \phi)$. Since the integrands are vector quantities the change in direction of the unit spherical vectors must be considered. This is most readily accomplished by expressing all vectors in rectangular coordinates. Noting that for the distant field

$$\rho\doteq r-a\cos\beta,$$

and

$$\cos\beta=\cos\theta\cos\theta'+\sin\theta\sin\theta'\cos(\phi-\phi');$$

and defining B_E and B_H by

$$B_E=B\frac{a^2}{4\pi r}e^{-jkr}h_n^{(2)}(ka)$$

$$B_H=B\frac{a^2}{j\omega\epsilon 4\pi r}e^{-jkr}\left[\frac{1}{a}-h_n^{(2)}(ka)+kh_n^{(2)\prime}(ka)\right],$$

so that $B_H/B_E=-Z_H$ at the spherical cap $r=a$; then the rectangular components of F and A are given by:

$$\begin{aligned} F_x&=-B_E\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}\sin m\phi'\\ &\qquad\times\sin^3\theta' P_n^{m\prime}(x')d\theta'd\phi',\\ F_y&=B_E\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}\{mP_n^m(x')\cos m\phi'\sin\phi'\\ &\quad+P_n^{m\prime}(x')\sin m\phi'\cos\phi'\sin^2\theta'\cos\theta'\}d\theta'd\phi',\\ F_z&=-B_E\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}\{mP_n^m(x')\cos m\phi'\cos\phi'\\ &\quad-P_n^{m\prime}(x')\sin m\phi'\sin\phi'\sin^2\theta'\cos\theta'\}d\theta'd\phi',\\ A_x&=B_H\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}mP_n^m(x')\\ &\qquad\times\cos m\phi'\sin\theta'd\theta'd\phi',\\ A_y&=-B_H\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}\\ &\quad\times\{mP_n^m(x')\cos m\phi'\cos\phi'\cos\theta'\\ &\quad-P_n^{m\prime}(x')\sin m\phi'\sin\phi'\sin^2\theta'\}d\theta'd\phi',\\ A_z&=-B_H\int_0^{\theta_0}\int_0^{2\pi}e^{jka\cos\beta}\\ &\quad\times\{mP_n^m(x')\cos m\phi'\sin\phi'\cos\theta'\\ &\quad+P_n^{m\prime}(x')\sin m\phi'\cos\phi'\sin^2\theta'\}d\theta'd\phi'. \end{aligned} \tag{11}$$

The azimuthal integrations are over an orthogonal interval and present no difficulty, but the colatitude integrations are not over an orthogonal range and the evaluation of these integrals does not seem possible, analytically, in their present form. For small flare angles approximations may be made which permit the evaluation of the integrals in a sufficiently converging series.

The horn is assumed excited in the dominant TE mode used by Southworth and King in an experimental study of the radiation patterns of conical horns with which the calculated patterns are compared. We take

$$\begin{aligned} \cos\beta&\doteq\cos\theta+\sin\theta\sin\theta'\cos(\phi-\phi')\\ P_n^1(\cos\theta')&\doteq\sin(\alpha\theta')P_n^1(\cos\theta_0)\\ P_n^{1\prime}(\cos\theta')&\doteq-\frac{\alpha\cos(\alpha\theta')}{\sin\theta'}P_n^1(\cos\theta_0), \end{aligned} \tag{12}$$

where $\alpha=\pi/2\theta_0$. The Legendre function approximation may be checked against tabulated values for integral n and is very good except for a small range where θ' is nearly equal to θ_0. The approximation in $\cos\beta$ involves neglect of the variation of $\cos\theta'$ over the aperture of

[11] A. E. H. Love, Phil. Trans. A197, 1 (1901).
[12] H. M. MacDonald, Proc. London Math. Soc. E10, 91 (1911).

TABLE I. Propagation coefficients in conical wave guide.

		E_θ component	H_ϕ component
$kr \gg 1$	γ	$1/r+jk$	$1/r+jk$
	λ_g	λ	λ
	v_p	c	c
$kr \ll 1$	γ	$\frac{n+1}{r}+jk\frac{\pi}{2}\frac{2n+1}{(n-\frac{1}{2})!(n+\frac{1}{2})!}\left(\frac{kr}{2}\right)^{2n}$	$\frac{n+2}{r}-jk\frac{\pi}{2}\frac{(2n+1)(n+1)}{n(n-\frac{1}{2})!(n+\frac{1}{2})!}\left(\frac{kr}{2}\right)^{2n}$
	λ_g	$\lambda\cdot\frac{2}{\pi}\frac{(n-\frac{1}{2})!(n+\frac{1}{2})!}{2n+1}\left(\frac{2}{kr}\right)^{2n}$	$\lambda\cdot\frac{2}{\pi}\frac{n(n-\frac{1}{2})!(n+\frac{1}{2})!}{(2n+1)(n+1)}\left(\frac{2}{kr}\right)^{2n}$
	v_p	$c\cdot\frac{2}{\pi}\frac{(n-\frac{1}{2})!(n+\frac{1}{2})!}{2n+1}\left(\frac{2}{kr}\right)^{2n}$	$c\cdot\frac{2}{\pi}\frac{n(n-\frac{1}{2})!(n+\frac{1}{2})!}{(2n+1)(n+1)}\left(\frac{2}{kr}\right)^{2n}$

the horn. The phase factor may now be written as

$$e^{jka\cos\beta}=e^{jka\cos\theta}\cdot e^{j\frac{1}{2}ka\sin\theta\sin(\theta'+\phi-\phi')}\cdot e^{j\frac{1}{2}ka\sin\theta\sin(\theta'-\phi+\phi')},$$

and the last two factors expanded in a Fourier-Bessel series of the form

$$e^{jx\sin y}=\sum_{p=-\infty}^{\infty}J_p(\gamma)e^{jpy}.$$

Setting

$$\gamma=\tfrac{1}{2}ka\sin\theta,$$

$$e^{jka\cos\beta}=e^{jka\cos\theta}\sum_{p=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}J_p(\gamma)J_s(\gamma)\cdot e^{j[(p+s)\theta'+(p-s)\phi-(p-s)\phi']}. \tag{13}$$

The neglect of the variation of $\cos\theta'$ over the mouth is thus seen to be a first order approximation in that terms of the second order and higher are neglected. It should be noted that the extent of the approximation is a function of the length, ka, of the horn and, for long horns, the neglected portion of the phase factor may have a considerable effect on the integrals, even though θ' varies over only a small range. The approximation is best for short horns and for co-latitudes far removed from the polar axis.

In the expansion, Eq. (13), the azimuthal coordinate ϕ' appears only in the trigometric form $e^{-j(p-s)\phi'}$ and this expansion when inserted in Eq. (11) is multiplied by other trigonometric functions of ϕ' and integrated over the orthogonal interval $(0, 2\pi)$. This will cause the vanishing of all terms of the double sum, Eq. (13), except those for which a definite constant relation exists between the indices p and s, and the double sum is reduced to a single sum over the index p. The vector potentials then become, from Eq. (11) with $m=1$,

$$F_x=-j\pi/2B_E P_n^1(x_0)e^{j(ka\cos\theta+\phi)}\times\sum_{p=-\infty}^{\infty}J_p(\gamma)J_{p-1}(\gamma)f_x(p,\theta_0),$$

$$F_y=-j\pi/2B_E P_n^1(x_0)e^{j(ka\cos\theta+2\phi)}\times\sum_{p=-\infty}^{\infty}J_p(\gamma)J_{p-2}(\gamma)f_y(p,\theta_0),$$

$$F_z=-\pi B_E P_n^1(x_0)e^{jka\cos\theta}\sum_{p=-\infty}^{\infty}\{\langle J_p(\gamma)\rangle_M^2 f_z(p,\theta_0)+1/2e^{j2\phi}J_p(\gamma)J_{p-2}(\gamma)f_y(p,\theta_0)\},$$

$$A_x=\pi B_H P_n^1(x_0)e^{j(ka\cos\theta+\phi)}\times\sum_{p=-\infty}^{\infty}J_p(\gamma)J_{p-1}(\gamma)a_x(p,\theta_0),$$

$$A_y=-\pi B_H P_n^1(x_0)e^{jka\cos\theta}\sum_{p=-\infty}^{\infty}\{\langle J_p(\gamma)\rangle_M^2 a_y(p,\theta_0)+1/2e^{j2\phi}J_p(\gamma)J_{p-2}(\gamma)a_z(p,\theta_0)\},$$

$$A_z=j\pi/2B_H P_n^1(x_0)e^{j(ka\cos\theta+2\phi)}\times\sum_{p=-\infty}^{\infty}J_p(\gamma)J_{p-2}(\gamma)a_x(p,\theta_0); \tag{14}$$

where the "f" and "a" functions are defined by

$$f_x(p,\theta_0)=(\alpha/2)\{g_1(2p+1,\theta_0)-2g_1(2p-1,\theta_0)+g_1(2p-3,\theta_0)\},$$

$$f_y(p,\theta_0)=-j(\alpha/4)\{g_1(2p,\theta_0)+g_2(2p-2,\theta_0)-g_1(2p-4,\theta_0)\},$$

$$f_z(p,\theta_0)=j(\alpha/4)\{g_1(2p+2,\theta_0)-g_2(2p,\theta_0)-g_1(2p-2,\theta_0)\},$$

$$a_x(p,\theta_0)=(\alpha/8)\{g_2(2p,\theta_0)-g_2(2p-2,\theta_0)\},$$

$$a_y(p,\theta_0)=-j(\alpha/2)\times\{\tfrac{1}{4}[g_2(2p+1,\theta_0)+g_2(2p-3,\theta_0)]-[g_1(2p+1,\theta_0)-g_1(2p-3,\theta_0)]\},$$

$$a_z(p,\theta_0)=-j(\alpha/2)\times\{\tfrac{1}{4}[g_2(2p-1,\theta_0)+g_2(2p-3,\theta_0)]+[g_1(2p-1,\theta_0)-g_1(2p-3,\theta_0)]\}, \tag{15}$$

and the "g" functions are given by

$$g_1(K,\theta_0)=\frac{\alpha e^{jK\theta_0}-jK}{K^2-\alpha^2}$$

$$g_2(K,\theta_0)=-\frac{4}{\alpha}\frac{Ke^{jK\theta_0}-j\alpha}{K^2-\alpha^2}. \tag{16}$$

The "g" functions and hence "f" and "a" converge as reciprocal summation index provided $|K|$ is greater

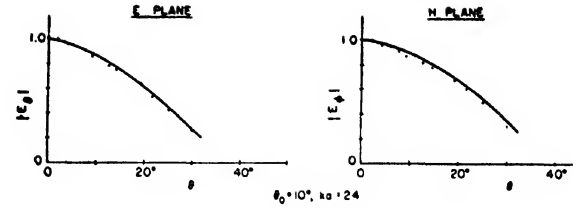


FIG. 2. Radiation pattern for conical horn of 10° flare angle. Solid curve is experimental and dotted points are calculated.

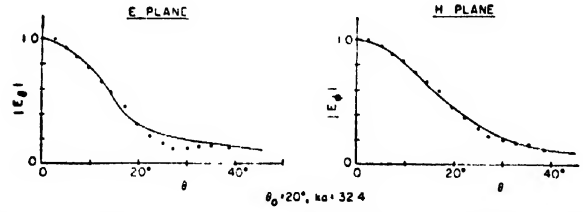


FIG. 3. Radiation pattern for conical horn of 20° flare angle. Solid curve is experimental and dotted points are calculated.

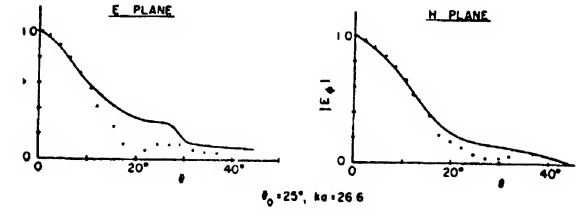


FIG. 4. Radiation pattern for conical horn of 25° flare angle. Solid curve is experimental and dotted points are calculated.

than α. The product of the two Bessel functions converges rapidly once the order p is greater than the argument γ. Hence the summations of Eq. (14) converge rapidly and, in general, only about γ terms must be included in the sum. These summations are over all integral values of p, positive and negative; however, the symmetry properties of the summands permits expression of the series as sums over only the positive integral indices. The only non-vanishing summations are those for which the total summand is an even function of the index of summation.

The vector potentials vary with radius as e^{-ikr}/r and thus if one ignores those radiation field components which vary faster than reciprocal radius, then Eq. (9) reduces to

$$E=k^2\frac{e^{-ikr}}{r}\{\theta_1[(\mu/\epsilon)^{\frac{1}{2}}F_\phi+A_\theta]-\phi_1[(\mu/\epsilon)^{\frac{1}{2}}F_\theta-A_\phi]\},$$

$$H=k^2\frac{e^{-ikr}}{r}\{\theta_1[F_\theta-(\epsilon/\mu)^{\frac{1}{2}}A_\phi]+\phi_1[F_\phi+(\epsilon/\mu)^{\frac{1}{2}}A_\theta]\},$$

where F and A are now the angular parts of the vector potentials, the radial parts being included in the e^{-ikr}/r factor. A receiving dipole oriented parallel to the E plane will then measure electric intensity patterns pro-

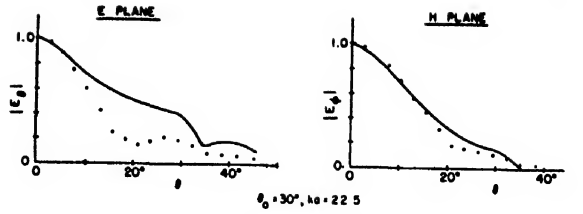


FIG. 5. Radiation pattern for conical horn of 30° flare angle. Solid curve is experimental and dotted points are calculated.

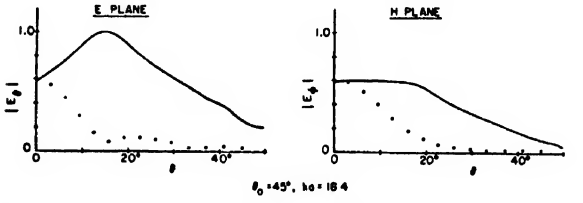


FIG. 6. Radiation pattern for conical horn of 45° flare angle. Solid curve is experimental and dotted points are calculated.

portional to

$$E_{\theta E}\propto F_z-A_x\sin\theta+A_y\cos\theta$$
$$E_{\phi H}\propto -F_x\sin\theta+F_z\cos\theta+A_y$$

in the $E(\phi=0)$ and $H(\phi=\pi/2)$ planes, respectively.

The "g," "f," and "a" functions have been tabulated for various flare angles up to 45 degrees. If γ is restricted to values less than eight, then ten terms of the series are adequate. The field point colatitude at which numerical information is obtained is $\theta=\arcsin 2\gamma/ka$ and will be a function of the length of the horn. In general, it is found that longer horns give sharper radiation patterns and the increased resolution and decreased maximum field angle is not undesirable. However, no information on the small scale radiation at large polar angles is predicted. The greatest contribution to the radiation arises from the F_z and A_y components; F_x and A_x are negligible in all horns up to about 30 degrees flare angle.

Radiation patterns were calculated for horn flare angles from 10 to 45 degrees and horn lengths, ka, from 18.4 to 32.4. Horn parameters were chosen to agree with those used by Southworth and King[7] in an experimental study of conical horns fed by cylindrical guide at 15.3 cm free space wave-length. The flare angles used by them are total flare angles and hence double the flare angle θ_0 used here; the lengths of their horns were determined from the diameter of the aperture D and the excitation wave-length λ by

$$ka=(\pi D)/(\lambda\sin\theta_0).$$

The resulting patterns are shown in Figs. 2 to 6 in which the magnitude of the electric field is plotted as a function of the field point polar angle in the E and H planes. The solid curves are replots of Southworth and King's Fig. 10,[7] while the dotted points are normalized values calculated by the assumed field method. Ampli-

tude patterns are given here so as to compare directly with the experimental patterns. Only one-half the pattern is shown, since the theory is completely symmetrical, with respect to co-latitude, about the polar axis. There is little difference in the experimental curves between symmetrical halves but, where such differences exist, the average of the two halves was used to draw the solid curves.

The agreement between theory and experiment is seen to be good in the E plane for the 10° and 20° horns, poor at 25° and 30°, and very poor for the 45° horn. In the H plane the agreement is excellent for 10° and 20° horns, fair for 25° and 30°, and very bad for the 45° horn. The complete failure of the theory for the 45° horn is perhaps not surprising in the light of the small angle approximation made. Certainly it is not sound to neglect the variation of $\cos\theta'$ when the flare angle is as large as $\pi/4$. The possibility existed that the large discrepancy in the case of this 45° horn might be due to a considerable third harmonic excitation at the mouth because of the large discontinuities at the feed junction and the relative shortness of this horn. However, a qualitative calculation of the third harmonic contribution was unable to account for this behavior. The theory, in its present form, seems unable to yield a radiation pattern whose maximum amplitude is anywhere but on the polar axis, whereas experiment shows that it is quite possible to obtain maximum radiation in other directions. This effect has also been observed by Chu and Barrow[13] in the sectoral horn even after great care had been taken to eliminate all higher order modes. This peaking off the polar axis seems definitely to be a characteristic of large flare angle horns with which the present small angle theory is unable to cope.

Another anomaly in the theory is its prediction of varying horn length, at constant flare angle, on the pattern. The horn length, ka, enters only through γ, the argument of the Bessel functions, and hence the only effect of a change in ka is to alter the field point polar angle at which the calculation yields numerical results. An increase in length will sharpen the whole pattern, but no change should occur in the relative amplitudes of the side lobes and other structure. This prediction is not wholly confirmed by experiment in that, while the directivity tends to increase with length, lobes become more prevalent. This discrepancy again arises from the phase factor approximation where, it will be recalled, the neglect of the variation of $\cos\theta'$ can be justified only for small flare angles and moderate horn lengths. From the figures it is seen that the E plane directivity, i.e., that polar angle at which the amplitude falls to .707 of its maximum amplitude, is always somewhat greater than that in the H plane, the relation being approximately $E_{\theta E}=E_{\phi H}\cos\theta$. However, the H plane pattern is usually much smoother, with less side lobe, and falls to zero somewhat faster.

In conclusion it is perhaps worth while to point out that for pyramidal horns the same field Eqs. (1) and (2) apply provided only that the associated Legendre function of the first kind $P_n^m(x)$ is replaced wherever it occurs by

$$L_n^m(x)=aP_n^m(x)+bQ_n^m(x).$$

The radial part of these conical horn solutions apply directly to the pyramidal case and hence the various propagation factors computed for conical wave guide follow identically for pyramidal guide.

[13] L. J. Chu and W. L. Barrow, Trans. A.I.E.E. 58, 333 (1939).

On the Diffraction of a Radar Wave by a Conducting Wedge*

R. B. WATSON AND C. W. HORTON
Defense Research Laboratory, The University of Texas, Austin, Texas
(Received February 14, 1950)

Diffraction patterns of radar waves have been measured about the edge of a perfectly conducting wedge. Theoretical patterns have been calculated using an asymptotic solution suggested by Pauli. Good agreement is observed between experimental and calculated patterns. The thin wedge tested showed much similarity in diffracting properties to a suitable semi-infinite conducting screen. The results may have application to diffraction effects about wedge-shaped wings on high speed aircraft and missiles.

A SOLUTION to the problem of the diffraction of electromagnetic waves by a perfectly conducting wedge was obtained by Sommerfeld[1] and extended by MacDonald[2] and Carslaw.[3] More recently, an asymptotic solution was developed by Pauli[4] which allows for relatively easy calculation of this diffraction. It was pointed out in an earlier article[5] (*I*) on the diffraction of radar waves by a semi infinite conducting screen that optical measurements showed deviations from the theory due to the nature of the screen material and screen geometry. Wolfssohn[6] suggests that these optical results might agree better with theoretical calculations for a wedge, thereby accounting for the thickness of the diffracting material and the shape of the sharpened diffracting edge.

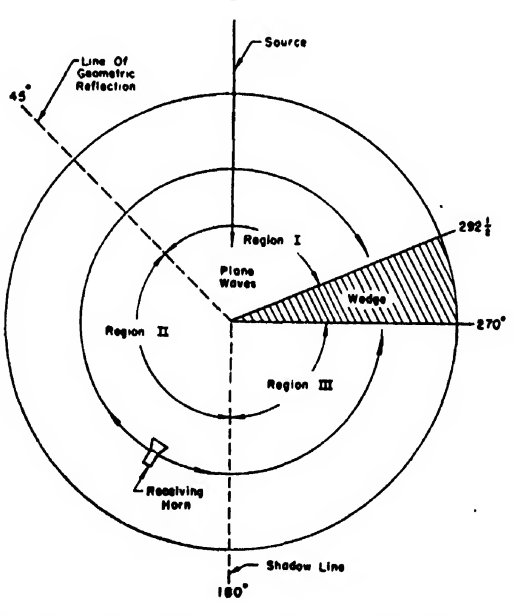


FIG. 1. Geometry of experimental arrangement.

The use of radar waves and metallic screens for obtaining diffraction measurements was justified fully in (*I*). The experimental arrangements described in (*I*) were used for the measurements on the wedge. The semi-infinite screen used in the earlier work was replaced by a hollow 22½° wedge made of two sheets of copper, each 30 in. × 60 in. × $\frac{1}{16}$ in. thick. These sheets were supported in the general manner described in (*I*) except for the use of dural channel members placed between the support members for the two sheets in such a way as to form the rigid 22½° wedge structure with the two ends of the sheets brought together to form a straight edge.

Observations of the incidental diffraction about the upper and lower edges of the wedge, the free ends of the two sheets, and various supporting members were made by taking measurements with the receiving horn at 15 in. and 30 in. radii and at two positions along the diffracting edge: at the midpoint and at the upper edge of the screen. The patterns taken at 30 in. radius showed considerable incidental diffraction effects from other parts than the major diffracting edge, in agreement with measurements in (*I*); thus, data are presented for the 15 in. radius only.

Figure 1 shows the geometrical arrangement of the diffracting wedge with respect to the source and receiver. The source was a sufficient distance from the wedge that the incident waves were effectively plane. The receiving horn was rotated in a circle about the diffracting edge and was thus oriented to face the edge of the wedge at all times and to give maximum sensitivity for waves diffracted from this edge.

Figure 2 shows a polar plot of the received signal when the incident wave was polarized with the electric vector parallel to the diffracting edge. Figure 2 also includes a theoretical plot for the same condition. The wave field was resolved for calculation into plane and cylindrical waves, representing the incident and reflected waves, and the diffracted wave respectively, for the regions I, II, and III. (See Fig. 1.) For these regions, the diffracted wave was calculated by means of the approximate formula given in Eq. (13) of Pauli's paper.[4] In the regions near the shadow line and the line

* The work described in this paper was done at the Defense Research Laboratory, under the sponsorship of the Bureau of Ordnance, Navy Department, Contract NOrd-9195.

[1] A. Sommerfeld, Math. Ann. 45, 263 (1894), and 47, 317 (1896) (general method of solution).
[2] H. M. MacDonald, *Electric Waves* (Cambridge University Press, London, 1902), p. 186 (specific form for the wedge).
[3] H. S. Carslaw, Proc. London Math. Soc. 18, 291 (1919).
[4] W. Pauli, Phys. Rev., 54, 924 (1938).
[5] C. W. Horton and R. B. Watson, J. App. Phys. 21, 16 (1950). This article will hereafter be denoted in the text by (*I*).
[6] *Handbuch der Physik*, Vol. 20, p. 278.

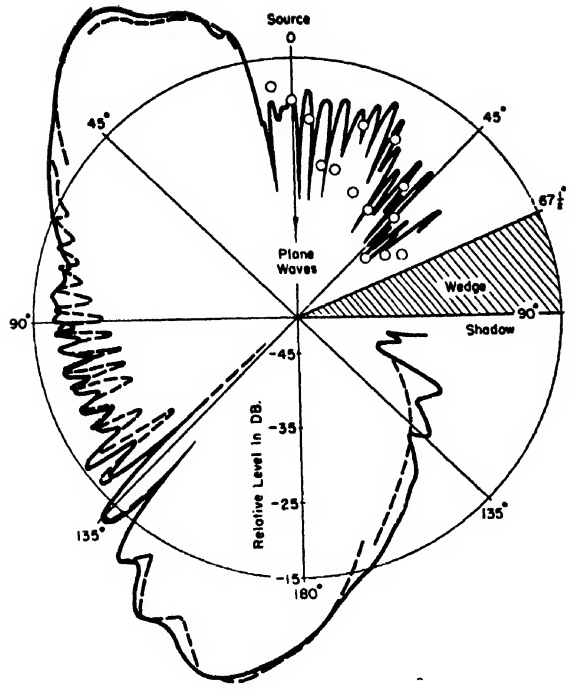


FIG. 2. Diffraction pattern for conducting wedge. E-vector parallel to diffracting edge. —— experiment; - - - and O theory.

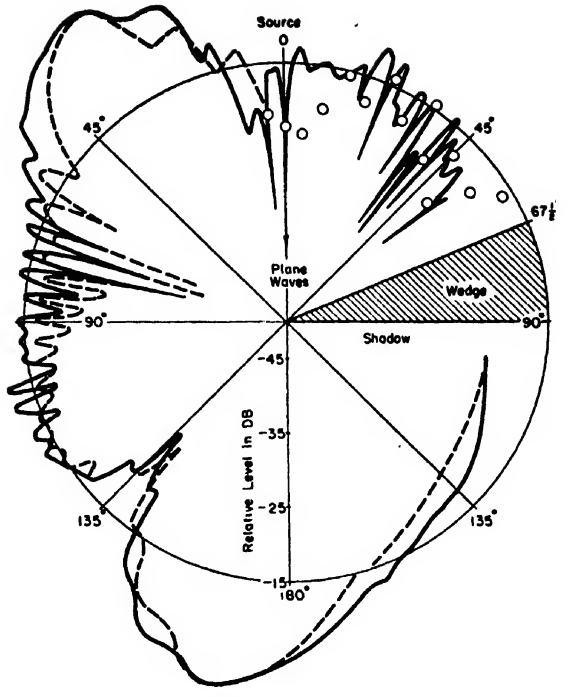


FIG. 3. Diffraction pattern for conducting wedge. E-vector perpendicular to diffracting edge. —— experiment; - - - and O theory.

of geometric reflection, the more accurate asymptotic series given as Eq. (35) in Pauli's paper was used.

Figure 3 gives a similar plot of the received signal when the incident wave was polarized with the electric vector perpendicular to the diffracting edge. Figure 3 further shows the computed pattern for this condition.

The agreement observed between theoretical and experimental patterns thus obtained again verifies the process of resolving the wave field into component fields. There appears a slight difference in the values for the two theoretical values for the angles removed 10° to 15° from the critical boundary lines. This difference arises because of the use of the two different types of approximate formulas in the various regions.

For comparison with optical results, the ratio of the amplitudes of the patterns for the electric vector parallel and perpendicular to the edge of the wedge has been plotted in Fig. 4 as a function of angle, for the shadow region only. The experimental curve agrees quite well with the theoretical curve in the region from 0° to 40°. Diffraction effects from other edges of the wedge and various surrounding objects become increasingly evident for larger angles. In line with Wolfssohn's suggestion that data taken for the wedge may fit optical data better than data taken for a semi-infinite screen, the optical data have been replotted here. There appears to be no significantly greater or less agreement than was observed in (I) for a semi-infinite screen.

The incident wave makes an angle of $-22\frac{1}{2}°$ with the normal to the forward sheet of the wedge. It is, therefore, of some interest to compare experimental measurements for the $22\frac{1}{2}°$ wedge with measurements made for a semi-infinite screen at an angle of incidence of $-22\frac{1}{2}°$. Such data are available in (I). An intercomparison shows only slight differences in detail. Similar agreement is obtained between experimental

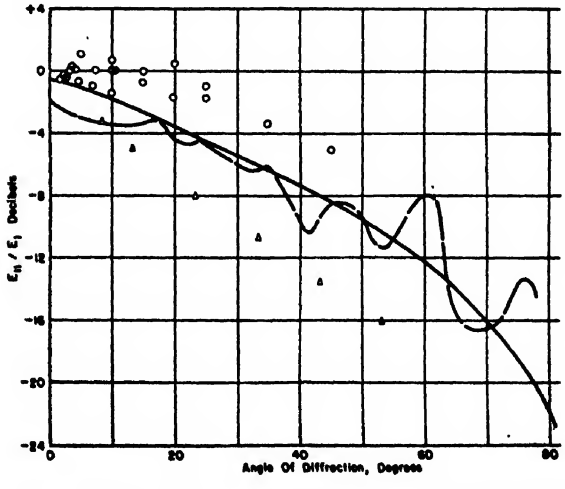


FIG. 4. Ratio of pattern amplitudes for electric vector parallel and perpendicular to diffracting edge. —— theory - - - experiment. Optical data: O Maey, Δ Jentzsch.

patterns for a quarter-infinite plane for an angle of incidence of $-22\frac{1}{2}°$ with the normal to the plane and patterns taken about the top edge of the $22\frac{1}{2}°$ wedge.

It is concluded that the diffraction caused by a thin wedge is not far different from the diffraction caused by the front surface of this wedge alone. Thus for such thin wedges, the rear sheet comprising the rear surface of the wedge is shown to be effectively shadowed by the front surface, and hence produces only small variations in the diffraction. These results may be of use in assessing diffraction about wedge-shaped wings on high speed aircraft and missiles.

Microscopic Observation of the Solidification of Small Metal Droplets

D. Turnbull and R. E. Cech
Research Laboratory, General Electric Company, Schenectady, New York
(Received March 6, 1950)

The solidification behavior of small metal (10 to 100 micron diameter) droplets has been observed on a high temperature microscope stage. An abrupt change in surface appearance and in the case of high melting metals a sudden brightening ("blick") accompanies solidification. The solidification temperatures observed for a collection of droplets may be widely distributed, but a significant, usually the major, fraction of the droplets supercool some maximum amount $(\Delta T_-)_{max}$ that is reproducible and characteristic of the metal. For many metals $(\Delta T_-)_{max} \approx 0.18$ times the absolute melting temperature. $(\Delta T_-)_{max}$ is not much changed by wide variations in the cooling rate and droplet size. The growth rate of metal crystals is very great so that the solidification rate of the droplets is controlled by the nucleation frequency.

INTRODUCTION

In another article[1] it was pointed out that if a metal specimen is broken up into a large number of isolated parts catalysts for crystal nucleation may be localized in a small fraction of the particles. Thus, in most of the particles crystal nucleation must take place homogeneously. In that investigation the solidification behavior of entire aggregates of particles was followed. An alternative procedure that is often advantageous is to remove a small number of particles from an aggregate and observe their solidification behavior microscopically. This is done by placing them in a stage recently designed by one of us[2] for microscopic observation at elevated temperatures. If only a small fraction of the particles contain nucleation catalysts there is a high probability that most of the ones selected for microscopic observation will be free of them. This principle has guided the present investigation.

A distinct advantage of the microscopic method of observation is its adaptability to the study of metals which melt at very high temperatures as well as medium and low melting point metals. Thus, it becomes possible by the use of one experimental technique to compare the solidification behavior of many metals.

For a long time[3] it has been known that small particles of liquid gold supercool significantly before solidification. A sudden brightening or "blick" occasioned by the liberation of the heat of fusion (recalescence) is known to accompany the solidification of the supercooled particle. This phenomenon is observed in the assaying of gold.

Mendenhall and Ingersoll[4] have observed this phenomenon in a number of other high melting metals including platinum, palladium, silicon, rhodium and iridium.

In their experiments metal particles 50 to 100 microns in diameter were supported on a Nernst glower in air or a CO_2 atmosphere, the solidification being observed microscopically. Particles of the same metal often solidified at widely different temperatures but the amount of supercooling, ΔT_-, was considerable and appeared to be roughly proportional to the melting point of the metal. However, the actual numbers for the maximum ΔT_- were given for only two of the metals, platinum and rhodium—370°C for both. These results of Mendenhall and Ingersoll were incidental to an investigation of another topic and their significance does not seem to have been realized.

In the current investigation the supercooling of many metals has been observed directly and the phenomenon studied in detail.

EXPERIMENTAL

In addition to its ease of adaptation to observation of metals having wide variety in melting temperature the microscopic method has the further advantage that particles isolated from one another on the hot stage do not require a film on their surface to prevent coalescence. It has been pointed out that[1] any crystalline protective film on the surface of a metal particle is undesirable in these experiments since it may catalyze crystal nucleation. On the other hand there are certain definite disadvantages of the method that must be recognized.

[1] D. Turnbull (to be published).
[2] R. E. Cech (to be published).
[3] A. D. Van Riemsdyk, Ann. Chim. Phys. 20, 66 (1880).
[4] C. E. Mendenhall and L. R. Ingersoll, Phil. Mag. 15, 205 (1908).

In producing aggregates of particles or in transferring them to the stage it is possible for nucleation catalysts to be picked up. Also, the particles, though free of crystalline films, must be supported on the stage and the possibility exists that the supporting surface or certain areas of it might catalyze crystal nucleation. In order to minimize this possibility, flakes of freshly blown quartz or pyrex glass were used as the supporting surface. These glass flakes, being amorphous, would not be likely to catalyze crystal nucleation. Another disadvantage of the microscopic method is that vibrations probably are more likely to promote solidification of small particles on a rigid supporting surface than particles in the interior of an aggregate. Sometimes these vibrations can be minimized by floating the particle in a suitable liquid flux.

MATERIALS

The source, purity, form, and method of preparation of particles are summarized in Table I. When the metal was supplied in the form of rod or wire small particles were made either by crushing with a mortar and pestle if brittle or by chipping with a sharp fragment of pyrex if ductile. Nickel particles and one sample of copper particles were formed from wire (fabricated from form listed in Table I) by means of a metallizing spray gun and the particles collected on a clean surface. These particles and those of other metals particularly susceptible to oxidation probably contained considerably more oxide than would be inferred from the purity listed. These metals were always given an *in situ* treatment with pure hydrogen in order to remove oxygen prior to observations of solidification. Different lots of some metals were obtained from two or three independent sources and the source designation used in the presentation of results is indicated in the table.

MICROSCOPE STAGE

Full details in the design of the microscope stage used in this investigation will be published in another paper.[2] A rough sketch is given in Fig. 1. The sample to be investigated was supported on a $\frac{1}{4}$-in. wide and 10 mil thick strip of molybdenum or platinum that was heated electrically.

Temperature was controlled by the amount of current passing through the ribbon heater and was varied by means of a variac transformer operated from a constant voltage source. By maintaining a constant power input into the heater it was possible to keep the temperature constant to $\pm 5°$ for long periods of time. The maximum temperature attainable was limited by the melting point of the heater metal. The stage is vacuum tight and its atmosphere (hydrogen, helium, or vacuum) is controlled by the indicated gas inlet. Temperature was calculated from the e.m.f. developed in an 0.005-in. diameter thermocouple welded to the underside of the heater. This e.m.f. was measured with a General Electric thermocouple potentiometer. The thermocouple was calibrated to read the temperature of the particles by accepting the known melting point of the metal as the standard. Deviation from the nominal temperature determined from standard calibration charts was usually +2 percent at the melting point. Solidification temperatures were corrected by assuming the deviation to be a linear function of temperature. A chromel-alumel thermocouple was used for temperatures below 1200°C and a platinum-platinum 10 percent rhodium thermocouple for temperatures greater than this. A new thermocouple was used for each specimen placed in the stage in order to prevent cumulative contamination from evaporation and condensation of the metals investigated. The error in the measured supercooling due to uncertainties in thermocouple and potentiometer calibrations and potentiometer readings is estimated to be ± 5 percent at most.

TABLE I. Source and purity of metals used in this investigation.

Metal	Source designation	Company	Purity (percent)	Form
Selenium		Johnson-Matthey	99.92	Shot
Bismuth	*A*	Eimer and Amend	99.999	Powder
	B	Johnson-Matthey	99.999	Rod
Lead	*A*	National Lead Co.	99.995	Ingot
	B	Eimer and Amend	99.95	Powder
Antimony		Johnson-Matthey	99.92	Ingot
Aluminum		Eimer and Amend	99.97	Powder
Germanium		Eagle-Picher Co.	99.99	Pellets
Silver		Handy and Harmon	99.97	Powder
Gold	*A*	Eimer and Amend	99.98	Powder
	B	Handy and Harmon	99.96	Wire
Copper		Amer. Smelting and Refining Co.	99.999	Rod
Manganese		Johnson-Matthey	99.99	Electrolytic sheet
Nickel		Int. Nickel Co.	99.98	Pellets
Cobalt		Johnson-Matthey	99.99	Powder
Iron		Nat'l Radiator Co.	99.99	Powder
Palladium		Amer. Platinum Works	99.95	Wire

PROCEDURE

Figure 2 shows the two types of experimental arrangements that proved most satisfactory. In (a) the heater has two bends or stirrups that were introduced to

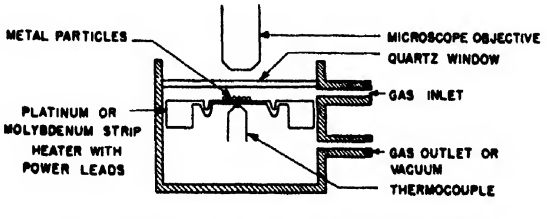


FIG. 1. Microscope stage with flat strip heater.

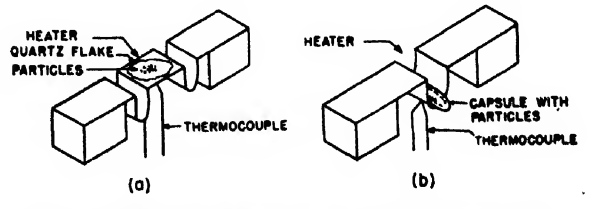


FIG. 2. Experimental arrangements used in investigation.

increase the effective length so that the temperature of the center horizontal section would be more uniform. This section supported a quartz flake on which a number of metal particles were placed for observation. In (b) a few metal particles were placed in the quartz capsule $\frac{1}{32}$-in. diameter by $\frac{1}{4}$-in. long which in turn was placed inside the single stirrup. A much larger number of particles could be observed at one time in arrangement (a) than in (b) in which the number was limited to 3–5. On the other hand, in (b) the temperature was very uniform along the length of the capsule and the calibration of the thermocouple more closely approximated that of the standard calibration charts. Also, (b) was much more satisfactory than (a) for observing volatile melts since in the latter arrangement the quartz window became fogged due to condensation of metal vapors on it. A new heater was used for each metal investigated in order to prevent contamination of the particles from condensation of metals accumulated in the heater during previous experiments.

An atmosphere of pure hydrogen, helium, or argon was maintained in the stage (excepting in the few instances that it was evacuated) during observations. Tank hydrogen was purified by passing it over a train consisting of a sulfuric acid bubbling tower and ascarite predryer, a copper chip furnace deoxidizer and a final dryer having dryerite and a liquid air trap. Tank helium or argon was forced through a purification train of two calcium chip furnaces in series operated at 550°C and 350°C.

In some of the experiments with metals melting at high temperatures Pyrex glass powder was mixed initially with the particles to provide a liquid coating during the solidification observations. It was thought that a liquid flux would cushion the particles against vibration and carry away or react with accidental crystal nucleation catalysts that might be present on the particle. However, it was demonstrated that the presence of a flux is not essential to the behavior to be described.

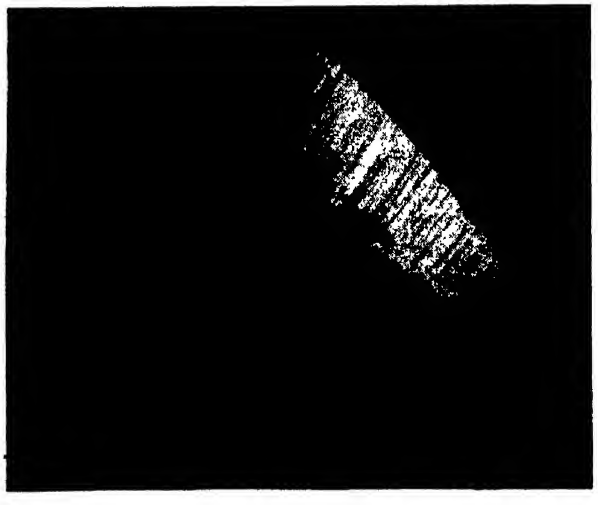


FIG. 3. Gold particles chipped from ingot. 25°C. 75×.

Melting and solidification of metals having melting points below 1000°C were inferred from the abrupt change of the surface condition of the particles that accompanied the processes. Upon initial melting, the particle lost its irregular shape and became spherical. Oblique illumination caused a bright spot or ribbon of light to be reflected from its mirror-like surface. When solidification occurred the rapid crystal growth in the particle caused a disruption of the smooth surface to one apparently having many sharp ridges and protrusions. These surface changes are illustrated in photomicrographs showing gold particles originally chipped from the ingot (Fig. 3), melted (Fig. 4), and solidified (Fig. 5). Melting and solidification temperatures were determined by association with the corresponding surface changes.

Solidification of metals having melting points greater than 1000°C was detected by the blick accompanying it. The initial melting of the particle permitted a thermocouple calibration as the particle became spherical. A further calibration was obtained by finding the maximum thermocouple e.m.f. that could be reached such that the particle did not blick upon cooling from the corresponding temperature. This e.m.f., which agreed within experimental error with that accompanying the initial shape change, was assumed to be that characteristic of the thermodynamic melting temperature. It was necessary to use filters for visual observations above 1100°C.

The melting and solidification temperatures of copper, gold, and germanium were measured by both the surface change and blick and the comparative results of the two methods are in complete agreement. Also, the maximum supercooling of bismuth particles (90°C) measured by observation of the surface changes is in close agreement with the value (90°C) determined dilatometrically by Turnbull.[1] In view of these facts there can be little doubt that the observed surface changes are caused by melting and solidification and that the blick observed in high temperature experiments is caused by solidification.

Particle diameters of the metals observed ranged from 10 to 100 microns excepting that some germanium particles were 400 microns in diameter. With the highest magnification that could be used, 240 diameters, it was not possible to detect the solidification temperature of particles smaller than 10 to 15 microns with certainty.

To determine the solidification temperature, the particles on the stage were cooled by decreasing the current through the strip heater. At intervals the cooling was interrupted and the thermocouple potentiometer exactly balanced and the e.m.f. recorded. The solidification temperature was taken to be that corresponding to the lowest exactly measured e.m.f. at which the particle was still liquid. Of the order of five minutes time usually elapsed in cooling from the melting point to the solidifi-

FIG. 4. Gold particles molten. 1070°C. 75×.

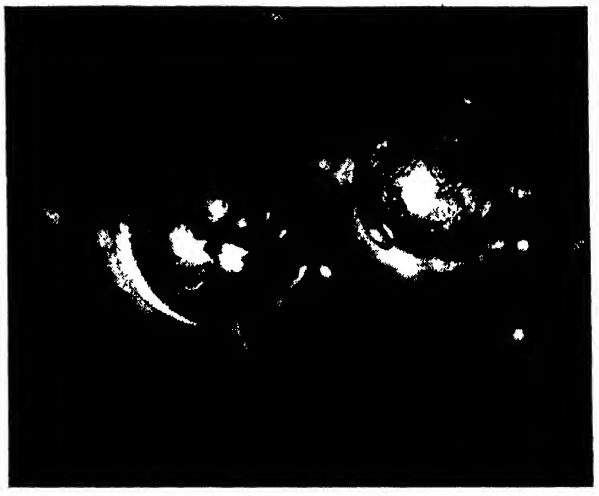


FIG. 5. Gold particles after solidification. 830°C. 75×.

cation temperature; most of this period was spent with the particles close to the solidification temperature. This cooling rate is rapid compared to that used in the dilatometric experiments but very small in relation to the maximum rate attainable (200°/sec.) with the apparatus. The temperatures of melting and solidification were checked several times for each particle or group of particles observed.

RESULTS

In general, the solidification temperatures of pure metal particles observed in a group ranged between the melting temperature, T_0, and some minimum temperature, $T_{\min}$. The results could be represented in terms of a number distribution curve relative to solidification temperatures or degrees of supercooling ΔT_-, where $(\Delta T_-)_{\max}=(T_0-T_{\min})$.

Depending upon the metal and circumstances (e.g., surface films) the number distribution of solidification temperatures sometimes had a maximum at $\Delta T_-\rightarrow 0$ (this is not certain for metals studied by the blick procedure since a blick cannot be certainly detected for $\Delta T_-<25$), or at $\Delta T_-\rightarrow(\Delta T_-)_{\max}$ or at some intermediate value of ΔT_-, and occasionally was fairly uniform so that no maximum was apparent.

It is an important fact that the solidification temperatures of a particular particle were not randomly distributed between $\Delta T_-=0$ and $(\Delta T_-)=(\Delta T_-)_{\max}$ throughout successive melting-solidification cycles but remained fixed for at least several and usually all such cycles. Occasionally the solidification temperature, after remaining nearly constant for several cycles, did shift abruptly to some other value and remained there in a series of succeeding cycles. This behavior is analogous to that of the one g continuous gallium samples described in a preceding paper.[1]

As noted before the sizes of the particles observed generally were between 10 and 100 microns. From a statistical standpoint the solidification temperature of smaller particles tended to be less than that of larger ones. However, it is significant that the solidification behavior of particular particles often deviated consistently from this generalization. Thus, 10-micron particles were occasionally observed to solidify at the melting temperature while 50-micron particles solidified at $\Delta T_-\approx(\Delta T_-)_{\max}$ throughout several melting-solidification cycles.

It appears reasonable to attribute the great differences in the solidification behavior of individual particles to the influence of minute quantities of accidental impurity particles. It is quite probable that colloidal particles suspended in the atmosphere and capable of serving as crystal nucleation catalysts become lodged in the quartz surface or on the surface of the metal particles during the assembly and operation of the microscope stage. (The experiments of Vonnegut[5] and Schaefer[6] have demonstrated that the concentration of colloidal material that can be effective in promoting the formation of snow is fantastically small.) Further, the abrupt change sometimes observed in the solidification behavior of particular particles can be attributed to the capture or dislodgement of a minute amount of impurity. *A priori* these impurities could either promote or inhibit solidification. However, there is much justification for the belief that such impurities promote crystal nucleation. This evidence has been summarized elsewhere[7] and will be discussed more fully in a following paper.[8] For the present purpose it will be assumed that minute impurities catalyze solidification. It follows that the only solidification temperature that could be characteristic of uncatalyzed solidification is $(\Delta T_-)_{\max}$. For this reason only the maximum supercooling $(\Delta T_-)_{\max}$

[5] B. Vonnegut, J. App. Phys. 18, 593 (1947).
[6] V. J. Schaefer, Chem. Rev. 44, 291 (1949).
[7] D. Turnbull, J. Chem. Phys. 18, 198 (1950).
[8] D. Turnbull (to be published).

TABLE II. Summary of results of microscopic observations on supercooling of small droplets.

Metal	Source	Melting point °K (T_0)	Method of observation	Atmosphere	Flux	Particle size (microns)	Maximum supercooling $(\Delta T_-)_{max}$	$(\Delta T_-)_{max}/T_0$
Selenium		493	Surface change	Hydrogen	None	75	25[a]	0.051
Bismuth	*A*	544	Surface change	Hydrogen	None	10–15	90	0.166
	B	544	Surface change	Hydrogen	None	20–50	90	
Lead	*A*	600.7	Surface change	Hydrogen or vacuum	None	15–50	67	0.151
	B	600.7	Surface change	Hydrogen	NaOH+KOH	10–20	69	
Antimony		903	Surface change	Hydrogen	None	15–30	135	0.150
Aluminum		931.7	Surface change	Hydrogen	None	50–100	48	
					NaOH	50–100	130	0.140
Germanium		1231.7	Surface change+blick	Helium	None	15	235	0.184
					Pyrex	400	219	
Silver		1233.7	Surface change	Helium	Pyrex	20–40	227	0.184
Gold	*A*	1336	Surface change+blick	Helium or vacuum	Pyrex	20–50	221	0.172
	B	1336	Surface change+blick	Helium or vacuum	None	40–50	230	
Copper		1356	Surface change+blick	Hydrogen or helium	None	15–50	236	0.174
Manganese		1493	Blick	Hydrogen	Pyrex	50	308	0.206
Nickel		1725	Blick	Helium	Pyrex	50–100	319	0.185
Cobalt		1763	Blick	Helium	None	20–50	330	0.187
Iron		1803	Blick	Helium	None	30–100	295	0.164
Palladium		1828	Blick	Helium	None	30–100	332	0.182

[a] Corresponds to maximum supercooling that could be reached such that crystallization did not occur on re-heating.

that could be reproduced under various conditions will be listed in the results.

The results are summarized in Table II. Values of $(\Delta T_-)_{max}$ determined in different atmospheres or in vacuum agreed within experimental error. The $(\Delta T_-)_{max}$ listed is the average value. Absolute melting temperatures of the metals listed in column 3 are from Kelley's[9] compilation.

FRACTION OF PARTICLES SUPERCOOLING MAXIMUM AMOUNT

In general the major fraction of metal particles melting at temperatures more than 950°C (Ge through Pd) did not solidify until supercooled to or nearly to $(\Delta T_-)_{max}$. Particles of source (*A*) gold constituted an interesting exception to this behavior. Without any Pyrex flux none of the particles were observed to supercool more than about 40°C. In the presence of the flux the supercooling was larger and a significant fraction supercooled to $(\Delta T_-)_{max}$. On the other hand the major fraction of gold particles chipped off gold wire (source *B*) supercooled to $(\Delta T_-)_{max}$ either with or without flux. The difference in the behavior of these two samples appears to confirm the principle set forth in an earlier publication[1] that crystal nucleation catalysts are more likely to be absent in metal particles freshly formed by breaking up a massive piece of metal than in particles formed "separately" or that have been permitted to accumulate catalytic dust on their surfaces for long periods. Another factor that might account for the difference is that the particles of source *A* gold actually observed were coalesced from several of the original particles which by themselves were too small for their solidification to be detected by the microscopic method. Thus, in effect more surface having the possibility of holding catalytic dust went into source *A* particles than into source *B* particles.

The distribution of particles of low melting metals with respect to solidification temperature had some interesting aspects and will be considered separately.

EFFECT OF ATMOSPHERE

Whenever a high melting metal having an oxide stable at the melting point was studied pure hydrogen was passed over the heated particles for some time and the melting and solidification temperatures were observed concurrently. After this the stage was flushed with pure helium and solidification observations were made in this atmosphere. When results were obtained for both atmospheres those for helium have been put in the table for the reasons that (1) hydrogen is significantly soluble in some of the metals and (2) e.m.f.'s of the platinum platinum-rhodium couple are not reproducible when the junction is in a hydrogen atmosphere. However, the maximum supercooling in the two atmospheres did not differ by as much as the uncertainty of the measurements.

It is significant that the supercooling of copper par-

[9] K. K. Kelley, Bureau of Mines, Bulletin No. 393 (1936).

ticles fabricated with a metallizing spray gun increased by about 100° after having been treated in hydrogen at 1000–1100°C for about 15 min. This may indicate that Cu_2O that formed in the fabrication process catalyzes the formation of copper crystals. Further evidence in favor of this view is that copper particles chipped from the original rod that supercooled the maximum amount in hydrogen or helium did not supercool significantly when the atmosphere was changed to air at low pressure.

The solidification of iron particles that had been hydrogen treated, evacuated, and sealed off in a quartz capsule also took place at the same temperature as in a helium atmosphere.

ALUMINUM

As received, the aluminum particles had a rod-like form that persisted in the liquid state due to the rigidity of the oxide skin. When placed on a quartz leaf with no flux practically all of these particles supercooled 48±2°C. The solidification range was so narrow that there was some temptation to assert that aluminum crystals nucleate spontaneously at $\Delta T_{-}=48$. However, in order to test the possibility that solidification was catalyzed by the oxide film, observations were made on solidification in a molten sodium hydroxide "flux." Under these conditions the particles spheroidized but a surface film (possibly sodium aluminate) remained. However, after heating 200° in excess of T_0, the particles supercooled 120° to 130°. This result appears to demonstrate that aluminum oxide does in fact catalyze the nucleation of aluminum crystals possibly by the mechanism suggested by C. S. Smith.[10] Of course there is no assurance that the film remaining in molten sodium hydroxide does not also catalyze solidification.

LEAD

Particles of lead placed on a Pyrex leaf in a hydrogen atmosphere solidified at temperatures ranging from the melting point to 60–65° below it at which only a minor fraction crystallized. As already noted, however, the behavior of particular particles was constant throughout successive melting-solidification cycles. When sodium hydroxide was used as a flux the fraction of particles solidifying at lower temperatures was greatly increased and a maximum supercooling of 69° was observed. On the other hand, in a flux consisting of a nitrate-nitrite mixture (56 percent KNO_3 and 44 percent $NaNO_3$) practically all the particles solidified at only 20–30° supercooling. A yellow film presumed to be lead oxide was present on the surface of the particles in the latter experiments. Thus, it seems likely that either lead oxide (PbO) or lead suboxide (Pb_2O) is a very effective catalyst for the crystallization of lead. Also it is interesting that the major fraction of source *A* particles supercooled to or nearly to $(\Delta T_{-})_{max}$ with no flux when the stage was evacuated to a pressure of 0.01 mm of mercury.

The maximum supercooling obtained in these experiments is about the same order as the supercooling of the major portion of lead in dilatometric and thermal experiments described in a preceding paper but is more than 10° less than the maximum measured in dilatometric experiments. Also, excepting when a sodium hydroxide flux was used, the fraction of source *B* particles that supercooled to 55–65° was much smaller in the microscopic experiments than in the thermal and dilatometric ones. These results perhaps may be accounted for by the fact that the particles observed in the microscopic experiments were considerably larger on the average than those used in the dilatometric and thermal experiments and so had more opportunity to contain or pick up nucleation catalysts.

BISMUTH

Although a major fraction of source *A* oxide coated bismuth particles supercooled 90° in dilatometric experiments only a minor fraction (presumed not to be oxide coated) supercooled the maximum of 90° on the microscope stage. However, it is significant that more than one-half of the particles made by breaking up source *B* bismuth rod in a mortar and pestle did not solidify unless supercooled about 90°. This difference in behavior of source *A* and *B* particles observed microscopically cannot be accounted for on the basis of different particle sizes for the *B* particles were significantly larger than the *A*. On the other hand, it seems probable that the difference in behavior of *A* particles in the dilatometric and microscopic experiments can be accounted for by the fact that many of the *A* particles observed on the stage had formed by coalescence of several small particles. As a result of this process, the proportion of particles containing nucleation catalysts was larger.

It is also significant that the temperature of solidification of bismuth particles that supercooled the maximum amount was not a function of the maximum temperature reached in the range of liquid stability when this was 10 to 100° greater than the melting point. This fact coupled with the lack of thermal history dependence in the dilatometric experiments proves that the thermal history effect observed in the solidification of large continuous bismuth samples is not an intrinsic property of bismuth and must be due to some extraneous effect.

The most important result of these experiments on bismuth is that the same maximum supercooling, 90°C, was obtained by two independent methods and with bismuth particles from two sources.

CRYSTAL GROWTH RATES

With the exception of selenium all of the droplets that had supercooled considerably solidified in a time too short to estimate, even to an order of magnitude,

[10] C. S. Smith, J. of Metals 1, 204 (1949).

after a nucleus had formed. From these observations it can only be maintained that the linear growth rates were greater than 1 mm/sec. It seems likely that crystals with relatively more complex structures such as bismuth and germanium would grow into their melts much more slowly than cubic crystals but this could not be proved or disproved by these experiments.

The extremely short period of growth relative to the total time required for solidification of a particle proves that "nucleation" is the rate-determining step in the reaction. Here the term nucleation is used in the broad sense as referring to a period of very slow growth of small crystals relative to large for whatever reason.

In no case excepting selenium was it possible to form a "glass" from the metal particles by quenching directly from the melt. Liquid selenium particles cooled to room temperature without crystallization. However, the glass particles so obtained crystallized upon reheating to temperatures 100°C or greater.* By cooling rapidly to some given temperature below the melting point and reheating, it was established that selenium nuclei were formed when the supercooling was greater than 25–30°C.

KINETICS OF ISOTHERMAL SOLIDIFICATION

Up until now it has been tacitly assumed that it is possible to assign "solidification temperatures" to particular particles. Since solidification is a rate process some justification of this assumption is necessary. As has been pointed out already, qualitative observations of Turnbull[1] have indicated that the temperature coefficient of the solidification rate of small metal particles is so very great that the rate changes by several orders of magnitudes within a narrow temperature range. Thus, solidification will be observed to take place only within this narrow range for widely different rates of cooling from the melting temperature.

This was confirmed by some observations on gold and copper particles. Liquid gold particles were on several occasions held at temperatures 180 to 190° below their melting point for one to three hour periods without solidification. They did not then solidify until cooled 230° below the melting point that corresponds to $(\Delta T_-)_{max}$ reported in Table II. Similar results were obtained with copper particles chipped from the rod.

DISCUSSION

The basic phenomena involved in the solidification of small metal particles now seem clear. In general, particles 100 microns or less in size can be supercooled to an extent far in excess of that ever attained on large continuous samples. Only for iron has a ΔT_- been reported[11] for a large continuous sample that is of the same order of magnitude (258°) as $(\Delta T_-)_{max}$ (295°) measured for small particles. For other metals having cubic crystal structures it has rarely been possible to supercool large continuous samples of the liquid more than 5–10°.

In an aggregate of metal particles the solidification behavior of individuals is quite specific and persistent but a significant fraction supercools some maximum amount $(\Delta T_-)_{max}$ that is reproducible and characteristic of the particular metal. The ratio of $(\Delta T_-)_{max}$ to the absolute melting temperature T_0 is 0.18 ± 0.02 (see Table II) for the metals studied having cubic crystal structures excepting lead and aluminum. (In making this calculation for iron and manganese, it was assumed that the crystal modification in equilibrium with the liquid at atmospheric pressure was nucleated although solidification in both instances takes place in a temperature range where a different crystalline modification is stable.) The lowest temperature of solidification of small particles is varied only a small extent with widely different rates of cooling from the melt and with changes in the particle volumes amounting to factors of 10^3.

The interpretation of these phenomena is to be given in a succeeding paper by one of us.[8]

ACKNOWLEDGMENT

The authors are pleased to acknowledge the assistance of Mrs. Joyce Rowland on many of the experiments herein reported. They also thank Drs. Leroy Apker and W. C. Dunlap for supplying samples of some of the metals.

* This agrees with prior observations of Apker and Dickey (private communication) on the crystallization of amorphous selenium films.

[11] Bardenheur and Bleckman, Stahl u. Eisen 61, 49 (1941).

Mass Spectrometric Study of Solids I. Preliminary Study of Sublimation Characteristics of Oxide Cathode Materials*

R. H. PLUMLEE AND L. P. SMITH
RCA Laboratories, Princeton, New Jersey
(Received March 6, 1950)

Mass spectrometric techniques have been utilized to study sublimation characteristics of materials used in electron tubes. Exploratory experiments and methods are described. Preliminary results concerning general processes induced by heat and by electron bombardment are discussed. Photographs are shown of mass spectrometer traces which demonstrate the effects described. Phenomena noted include (1) the inefficiency of electron bombardment for producing material transport from solids, (2) the thermal evaporation of alkali metal atoms and ions from bare nickel alloys, but no such evaporation from oxide cathodes formed on the same materials, (3) the complex thermal evaporation of oxide cathode constituents, (4) absence of positive ion evaporation from oxide cathodes, (5) the confirmation of mechanisms for negative ion formation in vacuum systems, and (6) the liberation of molecular oxygen from oxide cathodes which increases with electron emission drawn.

I. INTRODUCTION

ALTHOUGH considerable effort has been expended in the past on studies of structure-sensitive phenomena, the interpretation of the measurements produced has been inherently handicapped, for the most part, by a very deficient knowledge of the chemical constituents present in the materials studied. When the property under study, such as an electron emission process, is closely related to the functioning of the solid-vacuum interface as well as to the functioning of the interior of the solid and the solid-solid interfaces, the composition variable becomes quite complicated. The composition of surfaces is not only uncontrollable within the desired limits (because of diffusion exchanges with the interior of the solid, and because of absorption of foreign materials from the vacuum system), but it is also undeterminable by any currently known analytical technique. Surface analyses, ultimately of mono-molecular layers, made concomitantly with measurements of electronic properties, and without appreciably damaging the surface itself, would of course greatly facilitate interpretation of the measurements.

In developing an approach to the problem of surface analysis, mass spectrometric techniques are being applied to the study of materials of interest to the electronics industry. The results of these exploratory experiments described here are of interest in themselves, and, in addition, demonstrate the potentialities of a method applicable to a considerable number of problems in surface chemistry.

II. EXPERIMENTAL

The experiments were performed with a Nier-type spectrometer.[1] The tube was metal and glass with copper-gasket vacuum seals. The entire tube and liquid air trap were oven-baked at 450°C. A mercury diffusion pump was used. The spectrum was scanned by changing the ion energy. Recording was done by a carbon-paper printing mechanism from an obsolete facsimile receiver. The longest spectrum shown in succeeding figures represents less than five minutes of recording time. The ion accelerating potential was scanned linearly with time from high to low and vice versa at several fixed magnetic field values to cover the mass scale. The mass scale is thus inversely linear with ion energy, and the scales of some spectra are mirror images of others. The gaps in the strong peaks (e.g., Fig. 2A) are due to the recording arrangement whereby no printing was done during the scanning of a signal larger than the full ordinate scale (about 15 times the ordinate increment shown). Some of the experiments were recorded with an electron multiplier as detector. Most of the experiments were recorded with a conventional electrometer tube d.c. amplifier arrangement as detector. The ordinate increment shown for these traces represented usually an ion current of about 10^{-13} ampere. This has been translated into the estimated pressure sensitivity equivalent.

This sensitivity estimate was made from the measurement of the CO (mass 28) peak height and the associated ionization-gauge pressure reading during periods when the mass 28 peak constituted about 75 percent or more of the residual spectrum peak height. This estimate is necessarily rough because, besides uncertainties in the pressure measurement, it implies the same ionization cross section for all neutral particles. For many common gases, measured ionization efficiencies vary by less than a factor of ten[2] from that of CO. Although the ionization efficiencies of most of the species with which we are concerned have not been determined, this assumption of equal ionization cross sections is probably sufficiently accurate for present purposes. The necessity for maintaining the best possible vacuum discouraged the use of standard gases admitted for calibration purposes only. With solid samples being evaporated into the ionizing region, the sensitivity and optimum adjustment varied steadily so that frequent checks on the sensitivity were

* Presented in part at the meeting of the American Physical Society at Stanford University, December 30, 1949.

[1] A. O. Nier, Rev. Sci. Inst. 18, 398 (1947).

[2] John T. Tate and P. T. Smith, Phys. Rev. 39, 270 (1932).

essential. Using this calibrating method, fair agreement was found between the measured vapor pressure of nickel and literature values. For materials evaporated from the samples by the various methods employed, the measured pressures or vapor densities were also translated, by simple kinetic theory equations, into transport rates and into approximate times required for removal of mono-molecular layers.

The ion source most used was the double filament structure shown schematically in Fig. 1. It is essentially the usual gas analysis source with an added upper section for bombarding a heated sample. A similar design was used previously by Dart.[3] Bombardment may be done with positive ions (Kunsman source[4]) or with electrons. We have thus far used only electron bombardment. The sample bombarding region is separated from the ionizing region (analyzer) by a tungsten mesh. The structure therefore allows neutral particles evaporated from the sample to be analyzed in the usual gas analysis fashion, and charged particles from the sample region to be restrained, by suitable potentials, from entering the spectrometer analyzer; on the other hand, if so desired, charged particles from the sample may be accelerated through the spectrometer along with ions from the analyzing beam region, in which case two peaks for the same mass may appear. These peaks will be displaced from each other on the spectrum record by a distance corresponding to the potential difference between the two points of origin. Electron emission from

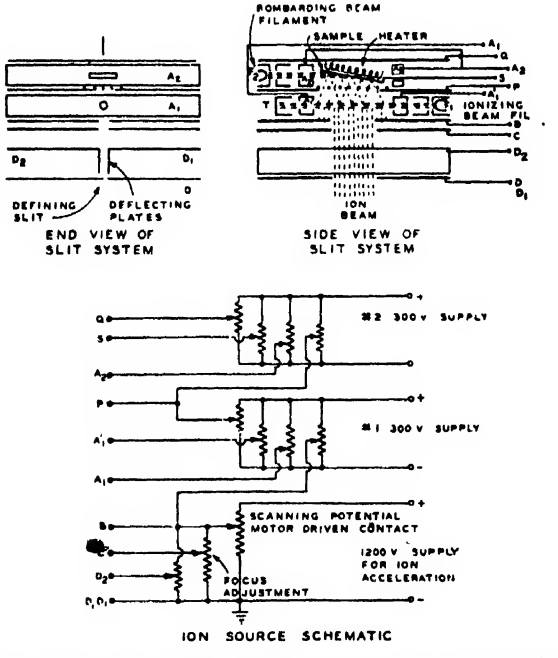


FIG. 1. Schematic drawing of an ion source and potential supplies used for electron bombardment study of solids.

[3] F. E. Dart, "The evaporation of zinc and zinc oxide under electron bombardment," doctoral thesis, Cornell University (1946); Phys. Rev. 78, 761 (1950).
[4] C. H. Kunsman, Science 62, 269 (1925).

the sample may be drawn to the surrounding box without traversing the mesh and without contributing to ionization in the analyzer region.

Both directly heated and indirectly heated samples were used, but mainly the latter. The sample was ordinarily in the form of a rectangular box (about $\frac{1}{2}$ in. long $\times\frac{1}{8}$ in. high $\times\frac{1}{8}$ in. wide) welded from a 0.002-in. sheet of the metal (or cathode base material) under study. A thermocouple junction of 0.0005 in. diameter Pt and Pt-Ir wires was welded to the sample, usually to the outside surface, centrally on the flat facing the spectrometer slit area. Oxide cathodes were made by spraying a carbonate mixture in a nitrocellulose-solvent vehicle over the box flat to be bombarded (the flat facing spectrometer slits). The sample was generally processed in the spectrometer. Processing was varied from experiment to experiment, but usually consisted of a brief operation, preliminary to bake-out, to check instrument performance, then a 450°C bake-out of several hours' duration with carbonate breakdown, the total pressure never rising above 10^{-3} mm Hg. After the carbonate breakdown, the liquid air was removed from the trap which was then torched empty of condensate, the liquid air was then replaced around the trap, and the bake-out continued for another hour or so. The residual pressure after bake-out was about 10^{-7} mm Hg with both source filaments operating and with the sample heated by its own heater to about 450°C. The source filaments were tantalum, coated on the emitting areas with tungsten powder and then thoria powder. They proved quite suitable for operations at pressures less than 10^{-5} mm Hg when no hydrogen was added to the residual gas of the system.

Figures 2A, 2B, 2C show typical binder decomposition spectra from an oxide cathode sample heated to 450°C before the bake-out. These materials, carbon-hydrogen-oxygen fragments of cellulose acetate and cellulose nitrate, were found to settle on all moderately cool elements of the ion tube and were not completely removed by the 450°C bake-out. They gradually distilled out of the ion source structure but could always be stirred up by moderately warming other portions of the ion tube.

Figure 3 shows the main residual spectrum at the end of the bake-out. The chief constituent of the residual gas was always CO. All residual peaks could be enhanced at any stage of operations by increasing the sample temperature, particularly in the case of oxide cathodes, indicating that an oxide cathode is never completely outgassed.

III. GENERAL ELECTRON BOMBARDMENT EFFECTS

Some cathode poisoning phenomena, well known to tube makers, have been attributed to dissociation or ionization of materials, adsorbed on grids or anodes, by electrons with energies from about 6 volts up to 25

volts.[5–7] A study of oxide decomposition using cathode poisoning as a method of detection was reported by Jacobs.[8] The technique of cleaning surfaces by electron bombardment has also been utilized.[9] Other experiments demonstrating material transport due to electron bombardment have also been described.[10] Dempster once employed electron bombardment of salts to generate ions for mass spectrometer studies.[11] Dart[3] found a small evaporation of Zn (and ZnO) from ZnO produced by electron bombardment.

It was expected, therefore, that electron bombardment might be the appropriate means of indiscriminantly stripping surface layers off solids for mass spectrum analysis; however, our electron bombardment studies of Ni, ZnO, Mo, Cu, and (BaSrCa) O surfaces have thus far found no general material transport effect that could definitely be separated from the purely thermal effect of the bombardment. No detectable evaporation of any sort was found to be produced by bombardment of these surfaces at low temperatures; however, at temperatures sufficiently high to produce a detectable thermal evaporation spectrum, addition of the bombardment, with the measured shell temperature fixed, enhanced the evaporation of non-metallic coatings only by slight amounts (less than a factor of ten, depending upon the bombardment). Electron bombardment energies up to 1400 ev and current densities up to 6 ma/cm² were used.

In no instance was an effect found which would indicate that surface layers of the bulk material of a solid are uniformly stripped away, at temperatures well below the sublimation point, by electron bombardment.

IV. SPECIFIC ELECTRON BOMBARDMENT EFFECTS

Although there may be numerous specific reactions between solids and low energy electrons, only one such was definitely established and one other probably established by these experiments. These are, respectively, the ejection of O^+ ions from (tank) hydrogen-fired molybdenum, and the ejection of Cl^+ ions from (tank) hydrogen-fired copper. The experimental arrangement by which the Cl^+ ions were obtained did not preclude the generation of these ions by negative ions (e.g., Cl^-) arising from the oxide cathode emission used for bombarding the copper. The ions were definitely formed at the metal surface and not in the gas phase.

Several times, indications were seen of O^+, OH^+, F^+, and Cl^+ liberation by bombardment of other metal surfaces, but the quantities were too small and transient for study. Circumstances suggest that they were all derived from adsorbed materials. Whether these materials were ionized on the surface before desorption, or ionized in the space adjacent to the surface was not operationally determinable.

The O^+ ions produced from the Mo surface had several unique properties, however, by which they differed from other ions in the spectrum. Figure 4 is a typical spectrum from that experiment. The ion source was the structure referred to earlier (Fig. 1) with the potentials $A_2-S=3$ v, $Q-A_2=0$ v, $S-P=10$ v, $P-B=2$ v, $A_1-B=1$ v, thereby transmitting positive ions from the sample region (superscript S peaks) as well as ions from the analyzer region (peaks without superscript S) to give a double spectrum. The displacement between peaks of the same mass should thus be expected to be

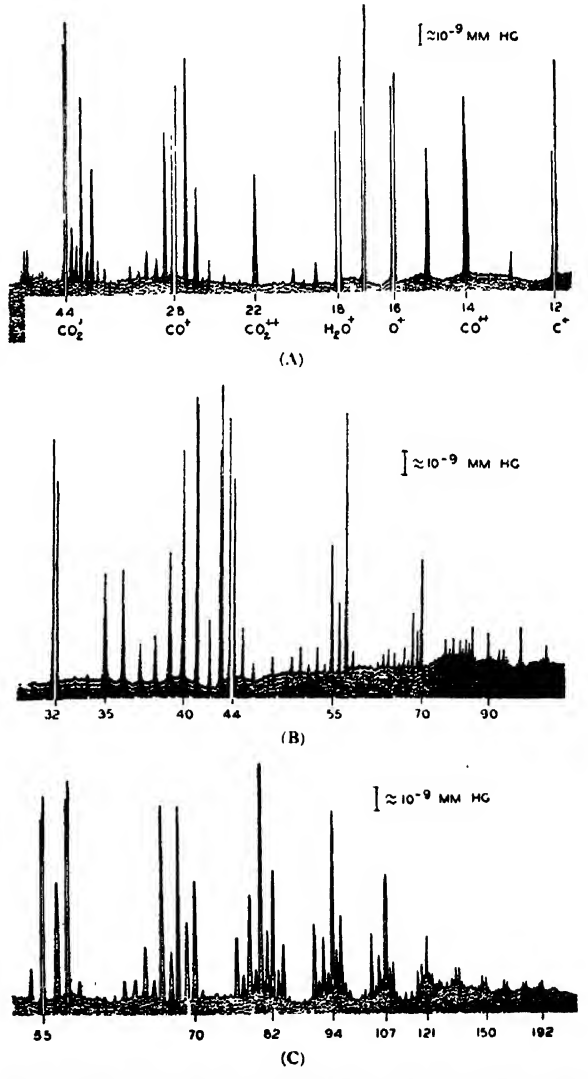


FIG. 2A, 2B, 2C, binder decomposition spectrum from oxide cathode sample at 450°C before bake-out.

[5] L. B. Headrick and E. A. Lederer, Phys. Rev. **50**, (A) 1094 (1936).
[6] Hamaker, Bruining, and Aten, Philips Res. Rep. **2**, 171 (1947).
[7] G. H. Metson, The Post Office Electrical Engineers Journal **41**, 204 (1949) (Part 4); Proc. Phys. Soc. (London) **62B**, 589 (1949).
[8] Harold Jacobs, J. App. Phys. **17**, 596 (1946).
[9] J. T. Randall, *The Diffraction of X-Rays and Electrons by Amorphous Solids, Liquids, and Gases* (Chapman and Hall, Ltd., London, 1934), p. 236.
[10] L. Jacob, Nature **157**, 586 (1946). G. I. Pokrovskii and S. T. Sinitsyn, J. Exp. Theor. Phys. (USSR) **8**, 1174 (1938). Chem. Abs. **33**, 7660[8] (1939). J. G. Trump and R. J. Van der Graaff, J. App. Phys. **18**, 327 (1947). I. Filosofo and A. Rostagni, Phys. Rev. **75**, 1269 (1949).
[11] A. J. Dempster, Phys. Rev. **11**, 316 (1918).

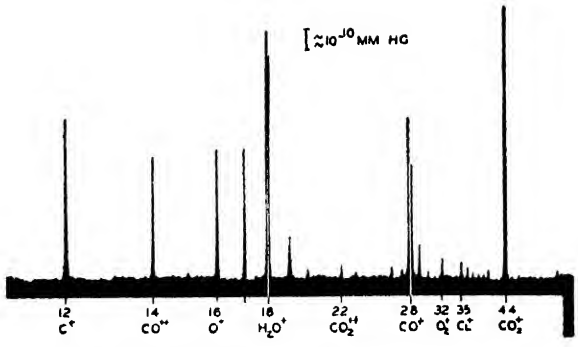


FIG. 3. Main portion of the residual spectrum after bake-out (and carbonate breakdown); sample temperature 450°C.

about 11 volts [the sum of $(S-P)$ and $\frac{1}{2}(P-B)$] corresponding to the potential difference in the two points of origin. Measurement of this displacement gives reasonable agreement for masses 18, 19, 23, 28, and 39, but not for the mass 16 peaks, for which the displacement is somewhat larger. Operation with a reversal of the $S-P$ field suppressed all the S peaks, less than 2 volts being required to suppress 18^s, 23^s, 28^s, 39^s, but about 11 volts being required to suppress 16^s completely. The excess kinetic energy of the O^+ ions (16^s peak) is the first property unique to those ions.

The second property of the 16^s peak not common to other S peaks was its decay under bombardment. The decay was roughly 50 percent in two minutes under bombardment of 0.9 ma per cm² with 300-v electrons and was independent of temperature. After an interlude of no bombardment, the peak could again be obtained at the original height upon resumption of bombardment. Even bombardment with 1000-v electrons provided a period of partial recovery from bombardment with 300-v electrons.

The third property of the 16^s peak was its temperature dependence. It could be obtained at temperatures above 300°C, and increased with temperature. It disappeared, when operations were suspended overnight, and could be obtained again only after a brief heating of the sample to about 900°C, a circumstance which suggests reaction of the Mo with residual gas of the vacuum system, or perhaps diffusion of some material from within the metal to give the surface coating required. The 16^s peak was produced with electrons of energy ranging from 40 volts to 1400, the highest used, with a broad maximum around 140 v. Below 140 v, the efficiency dropped gradually to extinction at 40 v. Since the remaining peaks (excepting the alkali metal peaks 23, 23^s, 39, 39^s, 41^s) showed almost no temperature dependence, they were concluded to be characteristic of the residual gas of the system. Sodium and potassium evaporated thermally from the sample as both atoms and ions. Upon application of bombardment, the increase in the 23^s and 39^s peaks due to electron impact ionization of metal atoms evaporating was small and the ratio of, for example, the regular 23 peak to this increase was not far from the ratios of $18/18^s$ or $28/28^s$ peaks. This established the relative transmission efficiencies for ions produced from the sublimate by the two beams. The $16/16^s$ ratio showed no relationship to this transmission ratio. Thus the 16^s ions must have originated at the surface and not in the vapor phase ionization of neutral particles evaporated from the surface. The excess kinetic energy indicates a dissociation from some excited state produced by bombardment of the surface layer.

V. THERMAL EVAPORATION FROM METALS

Studies of thermal evaporation from several† bare metals (cathode nickel alloys, and one molybdenum sample) showed, at moderate temperatures, copious evolution of alkali metal ions, and, at higher temperatures, evolution of alkali metal atoms as well as ions. At the higher temperatures, nickel alloys produced the expected evaporation of nickel atoms, and a rather profuse spectrum of carbon-hydrogen compounds which obscured the evaporation of minor constituents such as Mg, Mn, Ti, and Si. Molybdenum showed less carbon-hydrogen compounds. All samples had previously been fired in hydrogen.

The alkali metal evaporation was detected roughly as follows from nickel alloys: K^+ above 500°C, Na^+ above 600°C, K^0 and Na^0 above 850°C, the evaporation of all four species increasing with temperature, and the ratio of atoms to ions rising sharply with temperature. It may be significant that oxide cathodes made from the same nickel alloys showed no traces of alkali metal evaporation either as atoms or ions, at temperatures as high as 1100°C.

Referring again to Fig. 4, the 23^s, 39^s, and 41^s peaks are due to $^{23}Na^+$, $^{39}K^+$, and $^{41}K^+$ evaporation. These peaks appeared with both electron beams turned off, and showed temperature dependencies different from those of the regular 23, 39, and 41 peaks obtained with a retarding field ($P-S=5$ v) to prevent ions from the sample region entering the analyzer, and with only the ionizing electron beam turned on. On the other hand, the peaks for masses 12, 16, 17, 18, 18^s, 19, 19^s, 28, 28^s, and 44 were almost independent of sample temperature, and were thus identified as characteristic of the residual gas of the vacuum system. The total pressure of the system was less than 10^{-7} mm Hg according to ionization gauge readings.

VI. THERMAL EVAPORATION OF NEUTRAL PARTICLES FROM OXIDE CATHODES

The thermal evaporation spectrum from oxide cathodes depended very much on previous thermal history of the samples. In the early stages of the heat treatment (after carbonate breakdown and a "hot shot" at 1100°C to simulate activation) the evaporation from (BaSrCa) O cathodes was largely that of barium compounds, some

† A comparable thermal study of Cu was not made.

Ba atoms, and less Sr. The barium compounds evaporating were identified as BaO, and $BaCl_2$ (Fig. 5). (See Section XI.) The early evaporation, detected at temperatures as low as 600°C decayed at roughly 1 percent per minute and could be maintained at the minimum detectable level (about 10^{-10} mm Hg pressure in the ion source region) only by steadily increasing the sample temperature. During several hours of this procedure, the barium chloride evaporation rate declined to values small compared with the BaO evaporation. However, even though samples were held at temperatures around 800°C, for as long as 15 hours, the barium chloride could subsequently be made to register at sample temperatures of 1000°C to 1100°C. It is noted that the barium isotopes contribute fine structure to the peaks, and also that the Ba^+ peaks consist of contributions from ionization of both Ba and Ba compounds in the analyzer. It is also to be remembered that the evaporation is entirely uncharged, the ions being formed in the analyzer and not at the sample.

Prolonged operations with sample temperatures at 950°C±100°C developed, after several hours, a much slower decay rate for the Ba, BaO evaporation, of the order of a few percent per hour. The Ba^+/BaO^+ ratio was never larger than about 2/1 and declined slowly throughout. The ratio was definitely lower for the one platinum based cathode examined (Fig. 6) than for any of the nickel based cathodes formed from alloys containing medium or low quantities of reducing agents. The lowest ratio Ba^+/BaO^+ observed was about 0.4. The minimum peak ratio, Ba^+/BaO^+, is to be derived of course when no Ba evaporates, and is determined by the relative probabilities of the various decomposition processes with which ions are formed by the electron beam of the analyzer. Figure 7 shows even an appreciable Ba^{++} peak resulting from these processes.

The fact that some of the spectra shown were taken with the samples under bombardment is of no significance since the bombardment seemed merely to be an additional source of heat applied to the coating. Potentials were applied so that vapor-phase ions from the bombarding beam were not transmitted through the analyzer. In all cases, the absolute peak heights depended upon the past thermal history of the sample.

In general, the cathodes on nickel bases showed, in addition to the larger Ba/BaO ratio in the evaporation, appreciable Sr evaporation which was often nearly as large as the Ba evaporation (Fig. 7). At higher temperatures, a small Ca evaporation was detected also. The Sr evaporation decayed much less rapidly than the Ba and BaO evaporation.

Only traces of SrO evaporation were found, even at 1100°C, the highest temperature used. At this temperature, the literature[12] value of the vapor pressure of pure SrO is about 10^{-9} mm Hg, about ten times the minimum detectable by the instrument in some experiments. Considering the fact that high temperature effects were examined only toward the end of the experiments when the sensitivity may have diminished because of nonconducting coatings evaporated upon slit edges, the discrepancy between the literature value and the observed evaporation is not excessive. All samples were triple oxides, so that depletion loss of the more volatile BaO and SrO from outer layers of the coating would eventually leave an outer layer consisting very largely of CaO. At any stage, the vapor pressure of each constituent would be approximately equal to the product of its mole fraction and the vapor pressure of the pure material.

The cathodes examined were not what would be considered "well-formed," good emitters. This may have been due to contamination from the ion tube, or to improper activation. Emission values were usually about 25 ma d.c. per square centimeter for a cathode at 800°C.

[12] John P. Blewett, J. App. Phys. **10**, 668 (1939). A. Claassen and C. F. Veenemans, Zeits. f. Physik **80**, 342 (1933).

VII. POSITIVE ION EMISSION FROM OXIDE CATHODES

No positive ion emission was found from oxide cathodes at temperatures as high as 1100°C. The instrument was capable of detecting positive ion evaporation smaller than 10^{-10} amp. per cm^2. This does not conflict, of course, with literature references to positive ion emission from (BaSrCa)O coatings on tungsten at much higher temperatures.

VIII. THE NEGATIVE ION SPECTRUM

Examination of the negative ion spectrum agreed in general with the observations summarized by Sloane and Watt,[13] namely, that three principal mechanisms may be responsible for negative ion formation in partially evacuated systems, (1) thermal evaporation from surfaces, (2) conversion‡ of positive ions striking a surface to negative ions, which subsequently escape, often carrying part of the kinetic energy of the positive

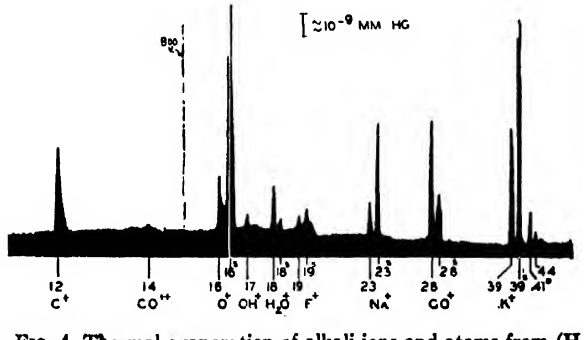


FIG. 4. Thermal evaporation of alkali ions and atoms from (H_2 fired) molybdenum at 1100°C, and ejection of O^+ ions by electron bombardment (4 ma/cm^2, 300 v) of molybdenum surface.

[13] R. H. Sloane and C. S. Watt, Proc. Phys. Soc. (London) **61**, 217 (1948).

‡ F. L. Arnot, Proc. Roy. Soc. (London) **158A**, 137 (1937) gives yield values, N_-/N_+, for this process ranging from 10^{-3} to 10^{-5} with 180-v positive ions.

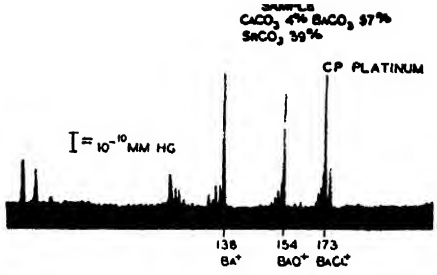


FIG. 5. Oxide cathode initial thermal evaporation at 800°C, all peak heights slightly exaggerated because of electron bombardment (4 ma/cm², 300 v).

ions, and (3) positive ion impact causing the dislodging of absorbed materials¶ as negative ions, which may also carry away some of the kinetic energy of the positive ions. These latter two processes were not distinguishable in the experiments here reported.

The operation demonstrated by Fig. 8 was performed with the source structure of Fig. 1, but with both electron beams turned off. Accordingly, with all source potentials negative with respect to ground and the ion collector, and with $P-S=100$ v, $B-P=2$ v, negative ions and electrons from the sample were accelerated through the spectrometer. In the spectrum, the displacement between peaks of the same mass number was found to be the voltage difference between the mesh P and sample S, and the non-superscript S peaks followed the potential of P. Therefore, these peaks were concluded to originate at the mesh while the S peaks were concluded to originate at the sample. With the field between S and P reversed so electrons and negative ions could not enter the analyzer region, no peaks were obtained. With $P-S=5$ v so electrons and negative ions were drawn from the sample to the mesh with low energy, then through the analyzer with high energy (several hundred volts), the non-superscript (mesh) peaks were again produced. This suggests that they were independent of the bombardment of the mesh by electrons or negative ions. Use of a magnet to deflect electrons (but not ions appreciably), so that the electrons did not accompany the ions through the analyzer, cut out all except the Cl^- ions. The interpretation is then that the large current of electrons in transit through the analyzer ionizes the residual gas to form positive ions which then follow the reverse path of the electrons striking slit edges, the mesh P, and finally the sample. On these surfaces, the incident positive ions generate the negative ions which are affected by the presence, in the source region, of the auxiliary magnetic field.

The largest negative ion peaks were those of the chlorine isotopes (Fig. 8) which were both temperature

¶ R. H. Sloane and R. Press, Proc. Roy. Soc. (London) 168A, 284 (1938). Their results, given for CO^-, Hg^+, CO^+, Hg^{++}, indicate current density yield ratios, N_-/N_+, for this process larger than 10^{-5}, the uncertainty arising from the measurement of the positive ion current. An unknown degree of surface coverage by adsorbed CO is also involved in this yield factor so the yield per impact per CO site is indeterminable.

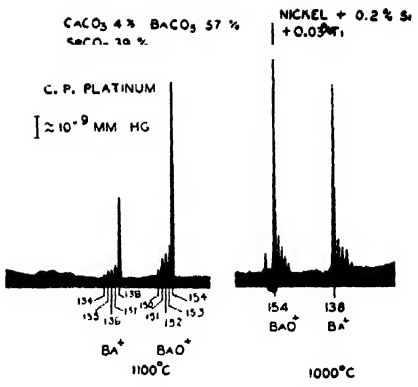


FIG. 6. Comparison of oxide cathode evaporation from platinum and nickel bases.

and pressure dependent, and were concluded to be largely of the thermal evaporation variety. They were found evaporated from an oxide cathode at temperatures as low as 600°C. The Cl^- peaks were much more sensitive to torching of the ion tube at various points than were the Cl^+ peaks formed in the usual operation of the analyzer. This observation suggests that the cathode collects neutral chlorine compounds or atoms and ejects the chlorine as negative ions. It also implies some storage mechanism whereby the Cl^- ejection need not be dependent upon the instantaneous vapor phase concentration of chlorine compounds as is the Cl^+ measured, but rather upon the concentration of chlorine accumulated in the cathode.

Masses 26^- and 14^- (Fig. 8, the superscript s peaks) were also generated at the cathode surface, but probably by one of the positive ion mechanisms, since there was little temperature dependence.

Mass 14^- was generated also, along with 12^-, at the unheated mesh partitioning the sample region from the analyzer section of the ion source (Fig. 1), by positive ion bombardment.

Trace quantities of still other negative ion masses were also found, such masses as 42, 68, 75, 130, but they were too small and erratic in appearance for study. They were apparently generated by a positive ion bombardment mechanism.

There was no visible correlation between the negative ion spectrum obtained as above and the usual residual gas spectrum (compare Fig. 3) obtainable a few minutes previously or subsequently (the change-over time from negative ion search to positive ion search was one to two minutes). In view of the variety of species inhabiting the vacuum system, any attempt to identify the composition of ions less obvious than C^-, $^{35}Cl^-$, $^{37}Cl^-$, seemed entirely speculative.

IX. EMISSION EFFECTS

Only one definite effect of electron emission from an oxide cathode was found reflected in the mass spectrum. This was an increase in the molecular oxygen peak

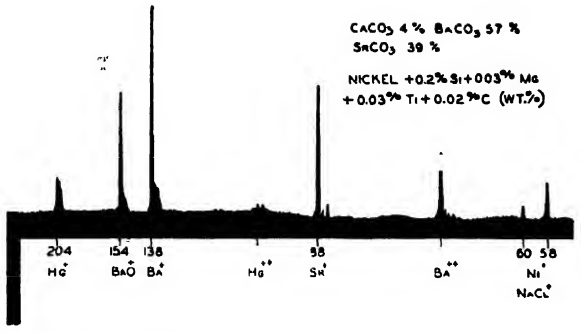


FIG. 7. Oxide cathode evaporation spectrum showing appreciable Sr evaporation at 1000°C.

(mass 32, Figs. 9, 10) as a function of emission drawn from the sample. The emission was drawn to the box surrounding the sample so it did not traverse the mesh P and contribute to the ionization in the analyzer. The bombarding filament was of course turned off during the experiment.

The total gas pressure was less than 10^{-5} mm Hg of which the O_2 partial pressure was probably never more than 1 percent. Examination of the negative ion spectrum showed no evolution of O^- or O_2^-. The evidence thus supports the conclusion that the oxygen left the cathode surface as O_2 molecules. The ratio, N_{O_2}/N_e, of oxygen molecules to electrons evaporated from the sample was estimated at about 10^{-5}.

X. CATHODE BREAKDOWN

The most significant fact of the breakdown was that it was never completed—each of the usual components (Fig. 11) seems to be evolved at a dwindling rate throughout the cathode life. The CO evolution, after the bulk of the decomposition was completed, was the most intense and by far the slowest to decay. The CO_2 and H_2O evolution usually assumed relatively small rates after several hours of sample operation at one particular temperature, only to be renewed at higher, slowly decaying, levels when sample operation was carried to a higher temperature. For nickel alloys and oxide coated samples, CO pressures less than an estimated 10^{-7} mm were never observed; for the Mo sample the CO pressure was about tenfold lower.

XI. DISCUSSION OF RESULTS

A. Electron Bombardment Experiments

The bombardment experiments indicate first that, if there is a general vaporization or dissociation characteristically produced by electron bombardment of solids, it is too small at low temperatures for detection by ordinary mass spectrometric methods. If, as has been proposed,[8] bombardment transport is dependent upon high temperatures to permit escape of excited fragments before de-excitation by other means, it may not be separable from ordinary thermally activated reactions and evaporations occurring at the same temperatures. All cases of simple vaporization or dissociation induced by electron bombardment must be carefully examined to be certain that they are not derived thermally from the temperature rise of the surface. Very few experiments reported in the literature demonstrating mass transfer induced by electron impact are above suspicion in this respect.

The second aspect of the electron bombardment experiments is the evidence that surface layers may undergo specific reactions. These may be expected to vary in efficiency from values approaching those of impact processes in free molecules down to zero, depending upon the particular nature of the layer under bombardment. Various properties of both the surface material and substrate, in addition to simple bond energies, may be critical.

Since these experiments are exploratory in nature and inherently lacking in precision, only a semiquantitative estimate of the results can be presented. The minimum detectable yield ratio, N_0/N_e, of neutral particles removed to bombarding electrons incident at the target, was about 10^{-6}. The minimum detectable yield ratio for positive ions ejected by electron bombardment was $N_+/N_e \cong 10^{-8}$ (because no subsequent ionization step was necessary).

In the case of O^+ ejection from the Mo surface, the 50 percent decay time (e.g., 2 minutes under 1 ma/cm², 300-ev bombardment) was temperature independent, but the peak height increased with temperature. The current density yield ratio, N_{O^+}/N_e, was about 10^{-6} at 1100°C and 10^{-8} at 300°C for 300-v electrons. Because of the constant decay time, it was concluded that the yield variation reflected a temperature dependence of the oxygen supply rather than a temperature dependence of the removal cross section. Assuming that the removable oxygen was adsorbed on the Mo surface, either as atoms or in some molecular form, one O atom at most per Mo site, and that half the supply of O sites was removed in 2 minutes of bombardment (the recovery time was about 5 minutes) and assuming a (100) Mo plane surface, 3.14A unit cell length,[14] the bombard-

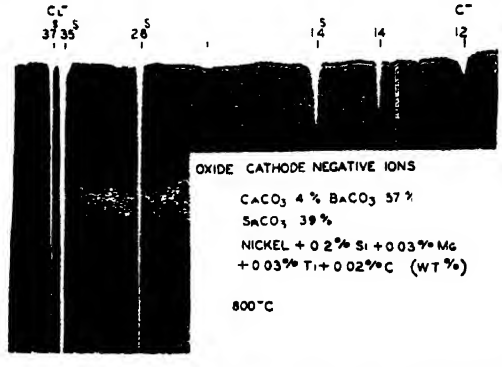


FIG. 8. Negative ions obtained from oxide cathode sample (* peaks) and from unheated tungsten mesh (mass peaks 12, 14).

[14] R. W. G. Wyckoff, *The Structure of Crystals* (The Chemical Catalog Company, Inc., New York, 1931), second edition.

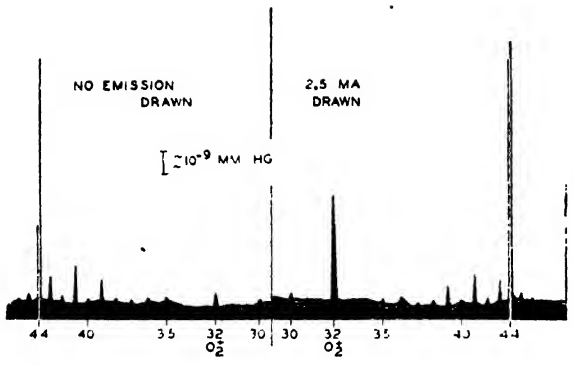


FIG. 9. Gas spectrum in the presence of oxide cathode sample showing increase in O_2^+ peak height with electron emission drawn from sample.

ment required for removing the adsorbed layer corresponded to about 10^3 electron impacts per surface atom site. Over-all yields of 10^{-6} and 10^{-8} thus corresponded to O site concentrations of one per 10^3 and 10^5 Mo sites, respectively.

Scant observations of the Cl^+ removal from Cu by bombardment with emission from an oxide cathode gave a single estimate of $N_{Cl^+}/N_e \cong 10^{-7}$. No decay was observed. If the Cl^+ resulted from conversion of Cl^- ions in the incident beam, the yield must have been near unity. This latter mechanism is considered rather unlikely in view of the fact that the Cl^+ ions were formed with essentially zero energy at the copper surface for bombarding beam energies up to several hundred volts. Had they been related to incident Cl^- ions in the beam, they would probably have carried away some of the kinetic energy of the incident ions.

It should perhaps be emphasized again that although yields as low as one neutral for 10^6 bombarding electrons or one ion for 10^8 electrons could have been detected, no such erosion of the bulk material of targets [e.g., Ni atoms from nickel sheet or Ba from (BaSrCa)O] was found, and the only removal observed was that of ions, formed from foreign materials adsorbed in some fashion on the target surface. Compared with reported positive ion (150 ev energy) sputtering yields[15] near 10^{-1} (one atom to ten ions) general electron sputtering yields are definitely trivial.

An application of these yield figures to the questions of target erosion and cathode poisoning is of interest. For the practical condition of electron emission (300 ma d.c./cm²) being drawn from a cathode to produce a general anode erosion yield of 10^{-6} (which could have been detected experimentally but which was not found), a maximum transport of 2×10^{12} neutral particles per second per cm² would be achieved.|| If this were all deposited on the cathode, one atom site in about 350 would be covered in one second, assuming a smooth (100) BaO plane of 5.50A[14] unit cell length for the cathode surface. In 1000 hours of cathode life, the anode erosion (e.g., copper) would amount to about 10^{-4} inch maximum. Thus anode erosion by this process would be insignificant, and sudden poisoning by the anode material would be unlikely.

Cathode poisoning of the low voltage variety[5–8] is usually associated with the drawing of electron emission to previously unbombarded tube elements and has been attributed to desorption, dissociation, or ionization of materials adsorbed on the target. Often beam energies too low to produce ionization are effective. The poisoning is usually observed to proceed to completion in less than one minute. A transport probably adequate for poisoning, about 0.1 monolayer cathode coverage per second, could be produced by the 10^{-3} efficiency observed for O^+ removal from Mo with space-charge-limited emission (30 ma/cm², 10 v, 0.5 mm spacing) drawn to a well-covered anode. This efficiency, computed as an effective collision cross section, is about 1 percent of peak values found for ionization of simple gases, hence higher values could occur. However, the gross removal efficiency at the critical anode potential may be indefinitely smaller than the peak value, because, depending upon cathode sensitivity to the poisoning agent, thermal energy distributions may determine the apparent threshold.

For pulsed emission operations (e.g., 30 amp./cm², 1 percent duty) the transport per second, aside from thermal effects, would be the same as for the 300 ma/cm² d.c. case; and for a 10^{-6} yield, in a single 100-micro-

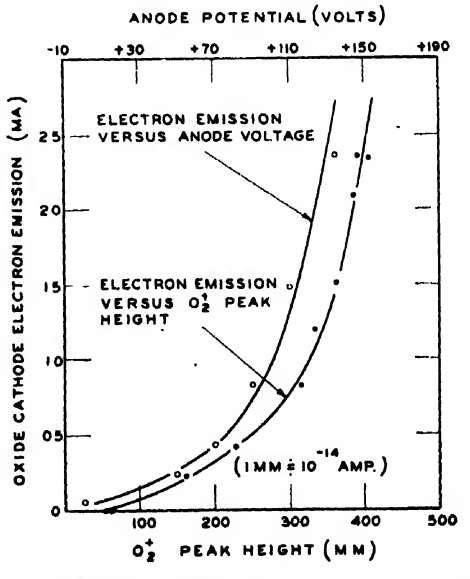


FIG. 10. Plot of electron emission drawn from oxide cathode (about 0.1 cm² coating area) *versus* O_2^+ peak height.

[15] K. H. Kingdon and Irving Langmuir, Phys. Rev. 22, 148 (1923).

|| For comparison, the intervening gas space maintained at a pressure of 10^{-6} mm Hg would contribute about 10^{14} incident gas particles per second per cm² of cathode, and would yield also about 10^{13} positive ions per second per mm of electron beam path (assuming 100-v electrons and a gas having the molecular weight and ionization cross section of CO). There would be only about 10^{10} molecules per cm³ of gas space.

second pulse, 1 percent of the d.c. transport for one second could accrue, 3×10^{-5} monolayer maximum. As indicated previously, about the same transport of positive ions might be derived from the gas space. Deactivation in a single 100 microsecond pulse as a result of reactions initiated by electron transit would thus hardly seem feasible. Higher yield electron bombardment processes, such as the ejection of adsorbed materials from well-covered anodes would appear essential to produce emission decay in this interval.

B. Evaporation Experiments

The evaporation of alkali metal impurity ions from other metals has been widely studied.[16] The re-evaporation of alkali metals as atoms and ions from heated filaments on which they had previously been deposited is also well known.[17] From these data it would be expected that, where impurity alkali ions are found evaporating from heated surfaces, atoms might be observed evaporating also. This has been reported[18] for Kunsman sources; however, we have found no reference reporting that impurity alkali atoms were detected evaporating from metals. The alkali metal atoms may conceivably have quite an important role in the reactions postulated to occur in oxide cathodes, since it was found that neither the alkali ions nor atoms evaporate from (BaSrCa)O coatings on metals which, when bare, gave a large evolution of alkali ions and atoms.

The evaporation of neutral entities from oxide cathodes was somewhat complicated. As anticipated from the accepted picture of the evaporation, the more volatile barium compounds were lost quite rapidly in the initial heat treatment and then much more slowly at a gradually diminishing rate. This agrees with the suggestion of Claassen and Veenemans[12] that the outer layers of individual crystals of (BaSrCa)O are quickly depleted of BaO and the rate shortly becomes limited by the diffusion rate.

The evidence of $BaCl_2$ evaporation from cathodes is based on the identification of peak groups (Fig. 5) as isotopic $BaCl^+$ (isotopic in both Ba and Cl). This identification is made with some hesitation in that no trace of $BaCl_2^+$ was found in the spectrum. These peaks were originally identified as BaO_2^+, $Ba(OH)_2^+$ from

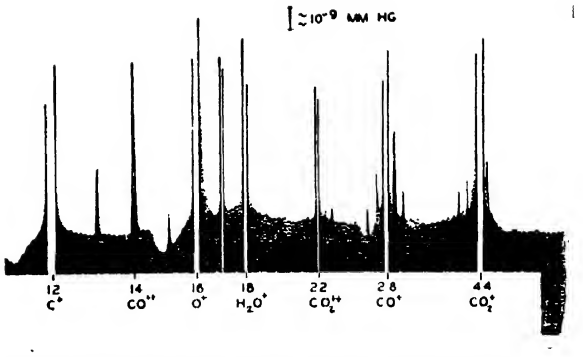


FIG. 11. Main portion of oxide cathode breakdown spectrum.

BaO_2 and $Ba(OH)_2$, respectively. Duplication of the result on several successive samples, together with the constancy of the isotopic (chlorine) ratios and better fit on the mass scale calibration favors the $BaCl^+$ identification. A $BaCl_2$ evaporation was reported by Hamaker, Bruining, and Aten.[6]

The Ba evaporation was expected, on the basis of the "free barium" picture and the mechanism of continuous core reduction of BaO, to be the most prominent feature of the cathode evaporation spectrum. The fact that it was always approximately equal to the BaO evaporation suggests it is not independent of the BaO evaporation, hence that it is perhaps limited by the rate of BaO reduction which in turn is limited by the rate of BaO diffusion to the reducing agent (and not by diffusion of reducing agent to the BaO). The presence of Sr in the sublimate, without SrO, argues against the possibility that the bulk of the Ba was oxidized by the residual gas as fast as it was generated and that the Ba^+ in the spectrum was all derived from ionization of BaO. The evaporation phenomena suggest that a considerable amount of information concerning the reaction mechanism could possibly be derived from a thorough study of evaporation rates and ratios and their decay characteristics.

XII. ACKNOWLEDGMENTS

Valuable assistance was derived from the suggestions of other members of the Physical Research Group, particularly Dr. D. O. North. Acknowledgment is made also of important contributions to the instrumentation by Dr. A. T. Forrester, Mr. W. B. Whalley, and Mr. J. F. Williamson, Jr., former members of the Laboratories staff. The recording system was designed by Dr. R. E Shrader.

[16] A. L. Reimann, *Thermionic Emission* (Chapman and Hall, Ltd., London, 1934).
[17] J. B. Taylor and I. Langmuir, Phys. Rev. 44, 423 (1933).
[18] A. M. Tyndall and C. F. Powell, Proc. Roy. Soc. (London) 136A, 145 (1932).

The Static Electrification of Dust Particles on Dispersion into a Cloud

W. B. KUNKEL*
University of California, Berkeley, California
(Received February 17, 1950)

An extensive study of the charge and size distribution of particles from 0.5 to 30 microns radius in dust clouds of diverse exemplary types dispersed in air under a variety of conditions ranging from blowing with minimum of turbulence to conditions of violent and maximum impact with various types of surfaces was made using adaptations of the Hopper and Laby modification of the oil drop experiment. It was observed that all dusts, including homogeneous dusts with no impacts on solid surfaces, were charged. In homogeneous systems the charges of opposite sign were equal, no *net* charge resulting, the number of particles of similar size with opposite charges being about the same. The magnitude of charges increased somewhat less rapidly than the surface of the particle. There was no correlation between size and sign of charge. There is strong evidence that charging occurs on separation of the contacts between particles in the dispersion of the cloud. Humidity did not affect the charging. Studies of heterogeneous systems making contact with solid walls of different composition from the powder gave consistent asymmetry of charge of varying degrees depending on the proportion of particles striking the surfaces relative to those just separated.

ALTHOUGH frictional electrification is the oldest manifestation of electricity known to man, it still remains today quite obscure as to the mechanisms active.[1] Not only are its mechanisms little known, but most of the data obtained are not reproducible and, in general, appear to be highly inconsistent and to a large extent even contradictory.

Most studies of frictional electrification antedate 1933 and involved studies using macroscopic systems. Work at this time, though stimulated by industrial processes affected by such electrification, appears to have been abandoned, primarily because with the many uncontrolled variables, the existing techniques were inadequate to yield consistent results. This is not surprising since most studies concerned themselves with net charges produced using particles of different sizes on carefully treated surfaces with controlled surroundings, temperature and pressure. As will appear evident later, the net charge analysis concealed some very important phenomena whose existence is useful in interpretation. The conclusions drawn were that charging depends on properties of the surfaces and involved the nature of the contact. Very small changes in these factors sometimes entailed large variations in the observed effects. The work of D. Debeau[2] in this laboratory showed the reproducible influence of absorbed gas films on surfaces of Ni, quartz, and rock salt, as well as water vapor exhaled from the walls. Recent studies of Peterson[3] showed similar changes of quartz surfaces by condensation of impurities exhaled from metal surfaces after evaporation.

That perhaps more information might be gleaned by studies of the individual charges on large numbers of microscopic particles with large surfaces relative to their masses, seems to have occurred to numerous workers. Thus, H. Sachsse,[4] in 1932, determined the charges on dust particles using the Millikan oil drop method. Similarly, it occurred to L. B. Loeb quite independently, for in 1933 he assigned to S. Chapman[5] the analysis of spray electrified particles by the same method. The method is too slow and tedious to obtain good statistics. Nevertheless, Sachsse found the charges to depend on the material as well as the methods by which dust clouds were generated. In most cases, positive and negative dust particles were present in equal numbers with only a small number of neutral particles. Using compressed air for dispersal, the average charge was found to be roughly proportional to the particle surface, but in most other cases, it seemed to be proportional to the radius. The observed charges ranged from 50 to 2500 electrons. In general, Chapman's results with atomized liquid particles confirmed these findings. The charges of both signs were present with few uncharged droplets for materials of high dielectric constant. A pure hydrocarbon gave 25 percent uncharged particles and low charges of both signs on the rest. The magnitude of charge was nearly proportional to drop radius and charges of up to 800 electrons were observed for the largest drops of 10-microns diameter. In 1937, N. Rosenblum[6] studied the behavior of charged dusts in an electrical field of alternating polarity. The amplitude of the photographed motion of the particles gave the charge, while the vertical motion gave the mass. The findings, in general, confirmed those of Sachsse.

With the contemporary intensified study of the solid state, the nature of solid surfaces thus derived may ultimately serve as a guide to experimental studies that will reveal much of the nature of the charge transfer at work. The improved modern techniques now available should also facilitate study, using microscopic systems so as to lead to consistent results such as Debeau observed. Currently, however, the opportunity offered since 1938 to study the charges and sizes of individual

* This work was supported by ONR Contract.
[1] L. B. Loeb, Science **102**, 573 (1945).
[2] D. Debeau, Phys. Rev. **66**, 9 (1944).
[3] J. W. Peterson, Phys. Rev. **78**, 91 (1950).
[4] H. Sachsse, Ann. d. Physik **14**, 396 (1932).
[5] S. Chapman, Physics **5**, 150 (1934).
[6] N. Rosenblum, Tech. Phys. USSR **4**, 564 (1937).

particles of a cloud of microscopic particles in a rapid fashion with good statistics seemed too good to pass over. Thus, as soon as the work of Hopper and Laby[7] on the evaluation of the electron using oil drops falling vertically in a horizontal electrical field and observed photographically by dark field illumination periodically interrupted by a rotating shutter was published, L. B. Loeb put students to the task of adapting this device for the study of dust clouds. The first attempt was interrupted by the war. In 1946, J. W. Hansen, an entomologist interested in the problem of insecticidal dusting, who had previously studied the electrification of such dusts using the older techniques, volunteered to set up the devices. Through his painstaking and skillful technique, the device was made to operate successfully and preliminary results were obtained.[8] Subsequently, the writer undertook to improve the device and with the improved techniques, which are described in a separate joint article by Hansen and the writer, the data here presented were obtained. As will be noted, the results obtained throw a considerable light on the mechanisms at work. There are still too many uncontrolled variables, as yet, to yield a complete picture of the processes at work, but certain mechanisms are definitely ruled out and other points clarified. The method has, of course, its limitations, but much more work with controlled atmospheres, as well as different systems, is still possible within the limitations encountered. Owing to the time involved in making the essential modifications for such work, this was not undertaken at this time. At a later date, it is possible that such studies will be undertaken.

EXPERIMENTAL TECHNIQUE AND LIMITATIONS

As indicated, the details of the apparatus and method of operation are elsewhere presented.[9] This paper will thus only concern itself with those features of the investigation affecting the results to be reported. As the studies do not require accurate evaluations of charges and particle magnitudes, certain inherent errors can be disregarded without correction, and measurements can be simplified in the interest of more data relative to accuracy of individual values.

Thus, uncertainties introduced by the irregular shapes of the particles and the consequent deviations from Stokes' Law were overlooked. The method, however, yields statistics for charge distributions for a whole cloud in relation to particle size, charge and numbers using random samples. While this method is capable of application to studies of carefully treated particles in well-controlled atmospheres, this was not attempted in the present study. Reasonable chemical and physical purification of the dusts was resorted to, and only dried room air was used. The dust clouds were dispersed by various means so as to extend from a gentle dispersion of the dusts with separation and a minimum of friction to more violent distribution with turbulence. Where needed, the blowing was controlled in as nearly a quantitative fashion as possible by means of an adjustable piston blow gun. The dispersion was also arranged so as to cover from a minimum of contact of the particles with surfaces to a maximum of such contact. Surfaces were as carefully prepared and controlled as contact with the room air permitted. Surfaces similar and different from the dusts, as well as mixtures of different dusts and pure dusts, were used. The measurements were checked as regards size distribution of the ultimate particles by microscopic examination of the dusts. This also served as a check on the absence of convection currents in the apparatus, which could seriously falsify both the statistics and charge sizes. For, if the size distribution as revealed by a microscopic study of the particles collected below the measuring field on a slide differs, especially in the region of small particles, from the distribution calculated from the rates of fall, there have been convection currents, either upward or downward, that produced the effect.

The principle of the method involved the production of the dusts by one of the various methods indicated below at the top of a settling tower of some meter in height (Fig. 1). At will, the cloud could be produced at different heights above the analyzing chamber. The cloud then settled down and along some small axial section fell through a slit into the observation or analyzing chamber containing a pair of vertical deflection plates, where it was photographed in a series of exposures on such a precalculated schedule to yield suitable sampling. The photographs indicated positive, negative and neutral particles in sizes ranging from 0.5

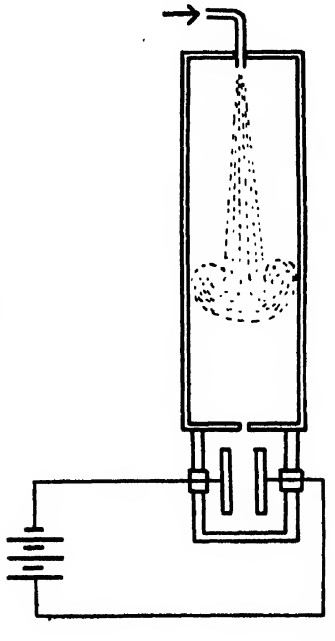


FIG. 1. Schematic drawing of apparatus showing principle of method.

[7] V. D. Hopper and T. H. Laby, Proc. Roy. Soc. A178, 243 (1941).
[8] J. W. Hansen, Phys. Rev. 72, 741 (1947).
[9] W. B. Kunkel and J. W. Hansen, Rev. Sci. Inst. 21, 308 (1950).

micron diameter to 30 microns. The charges varied from 0 to 30,000 electrons. The average charge per particle and the net charge of a sample depended on the substances used and processes of dispersion. By using large samples, these effects were reproducible. The cloud samples photographed in their measurements ranged from 200 to 2700 particles in a given cloud depending on the purpose of the investigation. The tracks of the particles observed consisted of a series of some ten dots separated by appropriate intervals, depending on the duration of the light flash and intervening dark time as set up on the rotating shutter in the optical parts (Fig. 2). Uncharged particles showed a vertical line of dots, unless the particle was largely irregular in shape. Particles out of focus by being too near the camera appear as circles; those too far away appear as stars. These appearances are predictable effects resulting from the optics. The tracks inclined to the one side or the other indicated charged particles of sign depending on the direction of the field used. The rate of vertical fall and the rate of horizontal movement allowed both charge and mass to be determined by insertion into the appropriate equations. In practice, the photographs were projected to a given scale of enlargement on a screen (usually 50×) and by means of a properly constructed nomographic scale, the charges

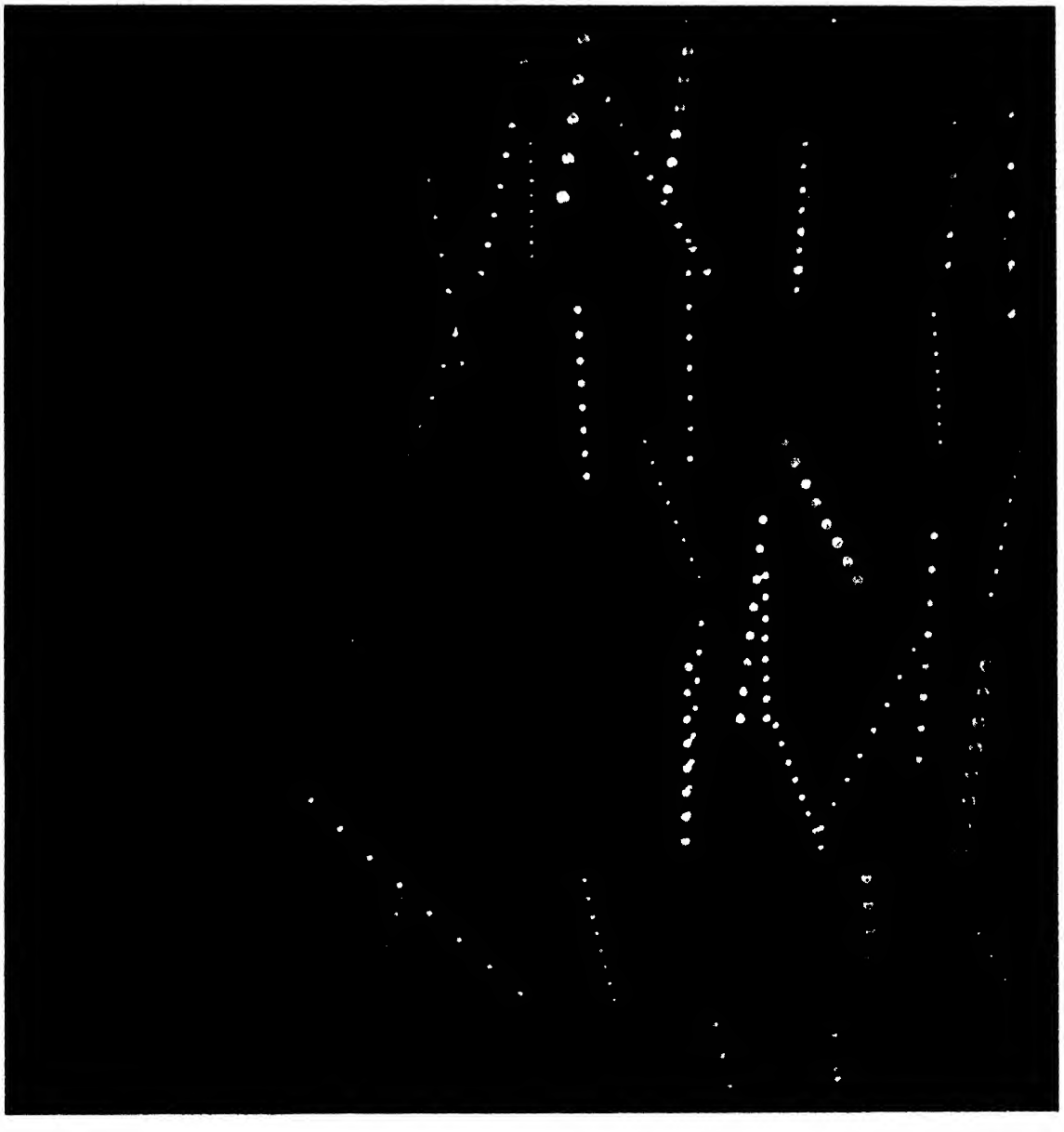


FIG. 2. Quartz particles settling in air under gravity in a horizontal electric field of 225 v cm^{-1}. 1 second exposure, 450 flashes per minute in dark field illumination. Total vertical distance shown is 5 cm.

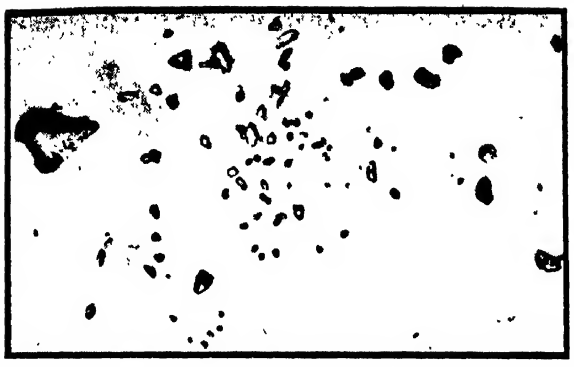


FIG. 3. Photo micrograph of quartz powder. Most particles are seen to be very little larger than one micron in diameter.

FIG. 4. Electron micrograph of quartz powder. Resolving power 50A approximately. The large particle was about four microns long.

and sizes could be inferred within a sufficient degree of accuracy for the purpose of statistics directly. When greater precision was required, more accurate data could be taken.

It must be realized that when a cloud is blown at the top of the settling chamber, there will be a segregation of particle size in falling through the chamber. Very shortly after the cloud is blown, the first very heavy particles reach the slit of the analyzer. As time goes on, the heavy particles settle out, and later—in fact even hours later—the smallest particles reach the analyzer. Thus the analysis of the cloud will consist of a properly timed and chosen set of photographs of the transient particles in the analyzer. To understand this, it is best to consider a characteristic movement.

Assuming a spatially homogeneous cloud to have been blown at a time $t=0$, the experiment consisted of a series of photographs of particle tracks taken at certain intervals. The number of particles observed in this fashion represented a certain fraction of the total number of particles in the narrow column settling through the slit at the top of the analyzing chamber. This fraction α, however, depends on the rate of settling of the particles, i.e., on the size, $\alpha=\alpha(d)$, where d gives the particle diameter in microns. In other words, the sample observed was not completely random but contained a certain size selection. Although we were not primarily interested in an exact representation of the size distribution in this study, some correction for this selectivity had occasionally to be made. A first correction could be applied by assigning proper statistical weights to the size groups observed on a picture. If W was the height of the region photographed and $v(d)$ was the rate of settling of the particles of size d, then the number of such particles included in the exposure time t is proportional to $W+vt$. The total number of such particles present in the column between successive exposures is proportional to vT where T is the average time between successive observations. Thus, the proper weighting factor for the particles of size d observed in exposures of duration t taken at intervals T is $\alpha=vT/(W+vt)$. If t and T are varied during the course of the experiment such that α for the average size observed on the pictures is kept constant, the sample taken did not falsify the actual size distribution greatly.

On the supposition that the above condition was fulfilled, the particle sizes and charges measured were tabulated and sometimes plotted. The charge-size distribution diagrams thus obtained for various samples give a good impression of the characteristics of the dust clouds in question. Frequently, the average charge per particle was calculated for successive size intervals, and the total charge present in the samples was evaluated. The results obtained and represented in this form were, in general, sufficient to provide all the desired information. In many cases, it was sufficient merely to compare the numbers of positive and negative particles in the samples and to estimate their charges only very roughly. In such a highly simplified representation of the experimental results, the element of chance introduced by the fact that only a sample of the cloud was measured could be expressed in a very simple manner. Since the particles were practically always charged and since the total clouds were always very large compared to the samples measured, the chance of finding a certain distribution in a random sample is given by the well-known binomial probability distribution law assuming homogeneity of the cloud. For large samples, such a binomial coefficient function exhibits an extremely narrow and steep maximum at the point of the actual distribution in the cloud. A half-width of such a maximum can be defined as that variation in sample distribution for which the probability is, say, one-tenth of the maximum value. If the samples are large enough, this half-width does not depend very much on the position of the maximum and is, therefore, evaluated most easily for the case of an equal distribution; i.e., for equal numbers of positive and negative particles in the cloud. This has been done for all experiments and will, henceforth, be called the standard deviation of the particular measured distribution.

The method of analysis outlined above is valid only if the cloud does not have its particles change in size or charge with time in the settling chamber. This subject

had been thoroughly investigated in all its aspects in view of some early results of Hansen.[10] The conclusions arrived at are as follows. In general, dust particles that settle quietly will stick together when they touch, thus, effectively growing less in number and larger in average diameter. A calculation has shown, however,[11] that for the cloud densities we were dealing with, the probability of such agglomeration was entirely negligible. We concluded, therefore, that the large number of aggregates invariably present in the clouds existed at least from the moment of dispersion on. This conclusion is confirmed by the fact that the amount of aggregation observed does not seem to be related to the length of the settling column.

Comparing the particle size distributions of the dust clouds for the cases of dispersion with much turbulence to those where the turbulence was slight, no significant difference could be discovered. From this result, it was concluded that the aggregates or glomerates had existed as such in the packed powders and were never separated into the ultimate particles by the stream of air, rather than that they were formed out of single particles during the violent agitation at the time of dispersion.

Another potential source of trouble is that the charges on the particles will, in general, tend to decrease in time. This arises from the presence of gaseous ions of both sign. The ions are constantly created in the air by stray radiation, and despite recombination, they will be maintained at a finite ion density in the air of the settling chamber. Highly charged dust particles suspended in the air will have an appreciable chance of catching these ionic neutralizing charges. In fact, a simple calculation reveals that for cloud densities of a thousand or more charged particles per cubic centimeter, almost all the ions created in air will attach to the dust before they have a chance to recombine. Thus,

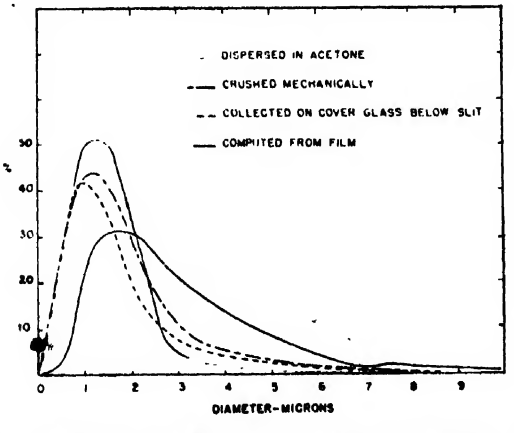


FIG. 5. Particle size distribution in precipitated sulphur. This graph shows the effect of a slight vertically downward convection in the analyzing chamber letting the particles photographed appear too large (settling faster than they should).

[10] J. W. Hansen, thesis, unpublished (1948).
[11] W. B. Kunkel, J. App. Phys. 19, 1053 (1948).

the total charge of one sign in a dust cloud will decrease linearly with time at a rate equal to the rate of ionization of the ambient air. As this rate is small, the decrease in charge is usually very slow, so that our experimental results will only be effected in the case of particles of very long settling times; i.e., for the very small particles. This is, indeed, what was observed. This problem has been investigated in detail and will form part of another paper. The results of the analysis allow of a proper computation of certain factors. Whenever an estimate of the initial charge of such extremely small particles is desired, an appropriate correction must thus be made. In addition, it should be borne in mind that the more highly charged particles will be neutralized more rapidly than those that were initially less charged. Thus, not only is the average charge of a cloud initially higher, but the initial spread of charges is larger. The corrections are not large compared to the charges involved and since little use will be made of the accurate evaluations, these corrections and their elaborate theoretical treatment will be omitted in this discussion.

THE POWDERS

The materials used in this study were chosen for a number of reasons. By selecting substances of widely varying physical and chemical properties, it was hoped, in the first place, to distinguish between the relatively important and unimportant variables limited, however, by three factors: the substance had to be easily obtainable in powder form of a grain size between 1 and 30 micron diameter; the powders had to be easily dispersible, with as little lumping as possible; and, lastly, the surfaces had to be reasonably clean. Unfortunately, this meant that neither hygroscopic nor chemically very active substances could be studied.

The desire for relative purity and chemical homogeneity, also, explains why no attempt was made to obtain and work with artificially rounded particles, rather than with the natural irregularly shaped substances.

The first part of this work logically undertook the study of the charges of as simple and elementary insulating substances as possible, in order to see what charges, if any, appeared on homogeneous systems. It was limited to two substances: quartz and sulphur only. These could easily be obtained in powder form of the desired grain size in fairly pure condition. The quartz was obtained by crushing and grinding. The sulphur was prepared by precipitation or crushing. The desired size range could be separated by sifting or, since neither of these are soluble in water, they could be sized by elutriation methods. Another advantage was that the containers could be made out of the same materials or in the case of sulphur, at least lined by the same material, so that sulphur only came in contact with sulphur and quartz with quartz. The main disadvantage with these substances was the particle shape, which, in general, was very far from spherical. Micrographs of

crystalline quartz are shown in Figs. 3 and 4. The electron microscope revealed that the surfaces frequently were far more irregular on a microscale than would be expected of crushed crystals. Although, in this sample most particles with diameters less than one micron had been carefully filtered out, a large number of invisible fragments were still present on the surface of the larger particle. They may have adhered to the larger particles on sizing, or they may have broken off the larger particles in the subsequent handling of the powder and then were later picked up. The presence of this coating of finer quartz dust in all observed particles must be borne in mind in the interpretation of the results. The sulphur particles were even worse in respect to regularity, with the additional drawback that they had a strong tendency to cling together very tenaciously, thus forming large aggregates of extremely irregular shapes. This could be ascribed to heavy electrostatic charging and small surface leakage. It was very difficult to break up these aggregates as shown in Fig. 5. Certainly, a gentle stream of air would be very inefficient in dispersing such a powder. This is characteristic of pure sulphur, so that in industrial practice, whenever a sulphur dust is desired, small admixtures of a suitable substance are added to prevent lumping. Since, initially, we were interested in the effect of sulphur on sulphur only, we could not make use of the dispersing admixtures. Thus, in our observations, the sulphur always contained a larger percentage of aggregate than the quartz. As indicated, the irregularity of the particles falsified the size and the charge as deduced by means of Stokes' Law,[12] and by sideways motion introduced apparent charges when none were present. To determine the magnitude of these simulated charges, test runs were made, photographing the dust as it settled under the influence of gravity alone. About 20 percent of the quartz particles showed slight deflections from the vertical, but the charges that would be calculated from these deflections in a usually fairly high field (about 200 v/cm) proved to be very small compared to charges actually appearing. Apparently, the constant tumbling motion in a field free space, due to Brownian rotation and differential heating of the settling dust particles, considerably reduced the amplitude of the helical motion,[13] which appeared to be much larger in the model study. There is evidence that in the presence of strong fields, these salutary rotations of the particles may have been absent. This came from the observation that the intensity of the light scattered by the particles fluctuated less when the electric field was present than otherwise. It is, thus, possible that, after all, the lateral drifts were somewhat larger in the actual charge measurements with the field. In the interpretation of the results, however, we shall consider these fictitious charges as negligible, inasmuch as they produce little error on multi-charged particles and, perhaps, gave a few more small lightly charged particles at the expense of neutrals than there were.

Another difficulty arose from the fact that both quartz and sulphur exist in different solid phases. It was, therefore, necessary to investigate in how far such a system could be regarded as homogeneous. The powder of precipitated sulphur was definitely rhombic, but when part of it was melted down to be used as a coating of the container, it would solidify in the amorphous phase. It was found desirable to allow at least two days to ensure that the coating had completely crystallized in the rhombic phase again. In the case of quartz, however, this procedure was not possible, since fused silica will not crystallize. We restricted ourselves, therefore, at first to crushed amorphous quartz powder in fused quartz containers, but later discovered that crystalline powder behaved in the same fashion. Either the surface of the carefully washed and thoroughly hydrated quartz powder stayed amorphous after dehydration, or the surface of the fused quartz assumed crystalline properties, or, possibly, in contrast to the sulphur, the difference between the two conditions was not sufficient to result in any detectable deviation from the homogeneous case. Consequently, the difference between amorphous and crystalline quartz was disregarded in subsequent work. Thus, the main study was done with a commercial grade of crushed quartz which was supposed to be quite pure. Since it might have been crushed in a metal mortar, it had to be washed in dilute hydrochloric acid and afterwards thoroughly rinsed in distilled water. After this treatment, the powder was dried for several hours at 120°C and stored in a closed glass jar. Nevertheless, it was found that the quartz would reach an equilibrium moisture content of a water layer of one to two molecules thickness surrounding the individual particles, which could only be driven off by prolonged heating to above 500°C. This dried powder would be so hygroscopic that it would acquire its former moisture layer very rapidly, even when placed in a dessicator containing phosphorous pentoxide. More surprising, however, was the fact that even in a very humid or almost saturated atmosphere, the water content of the powder, by actual observation, did not increase noticeably. Apparently, the average radius of curvature of these small particles was too small to permit water to condense on them in excess of the first layer that was adsorbed.

Similar observations were made on the sulphur powder, although the latter could, of course, not be baked out. An attempt to dry the sulphur at normal temperature in vacuum resulted in sublimation of sulphur, rather than in a removal of the last water layers. Only when the quartz and sulphur powders were placed in a supersaturated atmosphere could an increase in weight be detected; but in those cases, water condensed profusely on all surfaces, making the powders at once wet and lumpy. No intermediate stages of

[12] W. B. Kunkel, J. App. Phys. 19, 1056 (1948).
[13] E. E. Miller and Tanner, Phys. Rev. 77, 761 (1950).

moisture content seemed to be stable. Thus, the effect of water layers, if any, on the electrification of these non-hygroscopic substances could not be well-determined. It may be stated, however, that because of its relative stability, the monomolecular moisture film of such powders does not seem to be an important variable in the electrification phenomena. Certainly, the nature of these films rule out completely the old electrolytic ion exchange theory of tribo electricity.

The sulphur that was largely used in these studies was commercially produced, fairly pure precipitated powder, which happened to have just the right size distribution. Being a chemically fairly active substance, however, and moreover being produced in a chemical reaction, it was not expected to be completely free from impurities. Certainly, H_2SO_4 must have been present in spots, owing to slow oxidation in air. The insensitivity to moisture, however, indicates that these spots were few indeed.

Another substance which was chosen for its interest as a different type of powdered substance, e.g., a powdered carbohydrate, was a sample of very pure natural rice starch, the granules of which were very nearly spherical and uniform in size (between three and five microns in diameter). With this and some of the other powders mentioned, the studies had progressed so far that we no longer were interested in preventing the contact between dissimilar substances. Thus, it was possible to look at a wider range of powders which did not have to be made into coatings of the containers. Since it did not seem possible to take measurements on particles where the moisture had been completely removed and working in moisture-free systems without changing the experimental procedure, several runs were made with a talc powder, which was considered to be a naturally "dry" substance. The grain size and shape was very similar to that of the quartz powder.

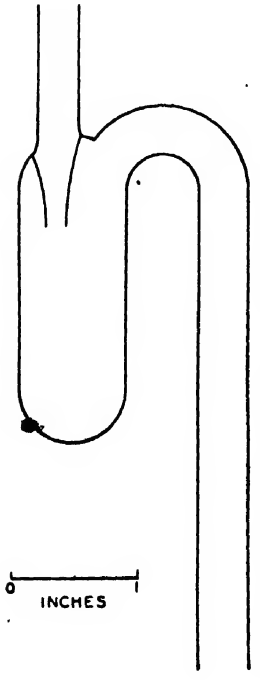


FIG. 6. Dusting tube made of quartz or Pyrex, sometimes lined with sulphur.

Obviously, among the simple elementary dusts, metallic powders come to mind. These are, however, very hard to produce in a satisfactory condition. In order to study the behavior of metallic particles, a nickel powder was prepared by rubbing a nickel rod with emery paper and separating the metal particles from the abrasive by means of a magnet. The powder had to be demagnetized and reduced in a stream of hydrogen. After this treatment, it appeared to be readily dispersible and of sufficiently fine grain size. The shape of these particles was, of course, extremely irregular, especially the larger ones, but since we had a nickel dust disperser, chosen because of its relative resistance to corrosion, it was considered desirable to investigate the behaviors of a powder of the same material. (It was only after the experiments had been discontinued that we were able to obtain a sample of the nearly spherical carbonyl iron powder. Such a powder, however, would be more heavily oxidized and, therefore, probably not representative of a metal surface.)

METHODS OF DISPERSION

In the interests of simplicity of operation, it was decided to employ dispersion by short blasts of air only. Since it was essential to the measurements that violent air disturbance in the settling tower be avoided, only a relatively small amount of air could be applied. For the same reason, more vigorous methods of cloud generation like explosions or mechanical shaking had to be discarded. We, thus, were rather limited in the range over which the factors involved could be varied. Nevertheless, as will be shown, this small range proved to be sufficient to permit the most essential conclusions.

In order to distinguish between some of the various mechanisms by which the charge separations might be effected, it seemed plausible, as a first step, to compare the result of a maximum of friction with that of negligible friction.

Hence, the simple dusting tube shown in Fig. 6 was designed so that even a gentle stream of air would cause strong turbulence, thus offering a maximum chance of frictional contact among the whipped up particles themselves and between the dust and the walls. Such tubes could easily be made of fused quartz or of Pyrex. One of the Pyrex tubes was carefully lined with sulphur. A similar duster was assembled out of nickel tubing. On the other extreme, it was impossible to disperse any of the powder without causing some slight friction between particles. However, it was attempted to reduce the contacts between particles in such dispersions after the initial separation to a minimum by blowing the powder very gently out of a small thimble-like container as indicated in Fig. 7. The motion in the clouds

thus produced indicated that turbulence was negligible in this case. The cups again could easily be made out of quartz or sulphur to avoid any possibility of contact between dissimilar substances when it was not desired. When, on the other hand, the effect of collisions of the dust particles with metal walls was to be studied, a third type of dusting tube was constructed especially designed to offer a large opportunity of particles striking the sides, Fig. 7. The number of baffles could be varied up to a maximum of fifteen per tube. The idea of forcing the stream of air into a zigzag path with sharp bends is fairly obvious and seemed to work well in many cases. The sample to be dispersed was placed on the second baffle from the top. Two of such tubes were constructed, one of nickel and one of platinum, avoiding all traces of foreign metals on the inside.

Early in the work, it was realized that the amount and velocity of the dispersing air stream had to be rather accurately controlled if reproducible results were expected. While the amount of charge appeared to be insensitive to the air blast, the amount of dust and thus the cloud density varied widely with the air stream used. Since these had to stay within certain limits to facilitate analysis and make statistical interpretation possible, it was clear that the air blasts had to be quantitatively reproducible. To achieve this, a blowing cylinder was devised, consisting of a brass tube of 6.25 cm diameter with a plunger bearing rubber "O" rings. This air gun had to be mounted upright above the dusting tube or cup. The outlet was at the bottom carrying a stopcock and a long thin detachable nozzle. The plunger could be lifted to entrap the desired amount of gas in the cylinder, the pressure could be regulated by loading the plunger with the proper weights. A further advantage of this design was that the blast to be applied did not necessarily have to consist of ordinary room air, but could consist of dried air or any gas which did not react with the tubing. This device performed very satisfactorily.

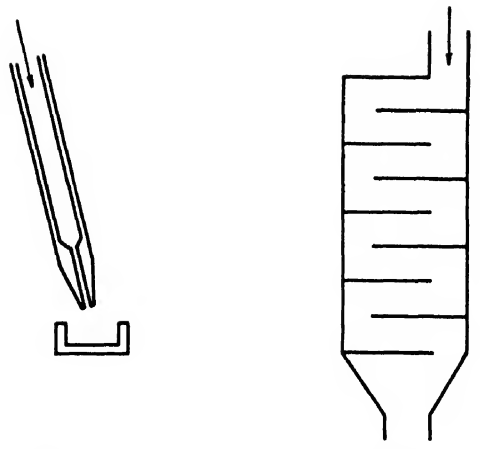


FIG. 7. Left: small cup and nozzle, made of quartz or sulphur. Right: metal duster made of nickel or platinum. The sample of powder to be dispersed was placed on the second baffle from the top.

THE RESULTS

All in all, more than eighty separate experiments were performed in this study, consisting together of more than 3000 exposures. Some 50,000 particles were photographed. The charge and size of not all of the particles were measured accurately. A large fraction of the charges and sizes were just estimated very roughly. The frequent repetitions of experiments under identical conditions showed the results to be completely reproducible. There is, therefore, little doubt that the findings here reported are reliable and of a general significance.

It must first be noted that every single cloud investigated *contained both positive and negative particles simultaneously*. Whenever any contact between unlike substances was completely avoided; e.g., blowing quartz powder out of a quartz tube or sulphur dust out of a sulphur cup, the numbers of positive and negative particles were about equal. Dividing the particles into groups of certain size ranges and calculating the average charge per particle for the various sizes, it was found that these were also equal for the two polarities. The average charge per particle increased, in general, somewhat more slowly than the square of the diameter, i.e., not quite linearly with the surface. This observation was quite general in all studies. Very few of the particles observed, in the range of sizes studied, were uncharged.

A good description of the charge, size and relative frequency distribution of the particles in such a dust cloud can be given graphically by representing each particle observed by a dot at the appropriate point in a charge-size plane (Fig. 8 for blown quartz dust). Since, in the particular experiment shown, more than 1500 particles were measured, only every fifth one was plotted in order not to overcrowd the graph. The averages drawn in, however, were computed from the total number. The picture indicated in Fig. 8 was typical of all of these homogeneous cases. Even the average charges never differed very much in these homogeneous studies, despite much variation in dispersion procedure. This is a rather remarkable result, since the amount of friction or the number of impacts between the particles presumably varied appreciably, depending on the amount of turbulence during dispersion. Particular values for fine quartz dust blown out of a quartz duster and out of a small cup, representing the extremes studied, are given in Table I. Similar values were found for sulphur dust. Several explanations of this unexpected behavior suggested themselves. The processes by which charge separation occurred apparently were not influenced by the difference in turbulence. This probably means that the charge separation occurred at the first separation of the contiguous particles and the charges only possibly redistributed themselves slightly at subsequent collisions which would not affect their magnitude appreciably. Whether these charges were

TABLE I.

SiO_2 Size Micron diameter	SiO_2 duster Number of particles +	0[a]	−	Average charge +	−	SiO_2 cup Number of particles +	0[a]	−	Average charge +	−
0–1	147	4	161	15.3	18.2	158	0	157	13.9	14.1
1–2	446	32	474	22.2	23.2	340	8	409	23.4	25.2
2–3	84	5	96	45.8	49.3	127	6	125	52.8	59.8
3–4	29	3	30	72.4	67.3	36	1	35	103.8	86.8
4–5	7	0	8	112	99.2	21	3	23	134	172
5–6	5	0	4	105	117	15	1	9	155	159
6–7	0	0	2		342	7	2	6	141	381
	Total numbers			Total charges		Total numbers			Total charges	
	718	44	775	20,000	23,000	704	21	764	27,000	31,300

[a] Here particles were considered neutral when their charges were too small to be determined; i.e. $q < 3$ for $d \approx 1\mu$ increasing to $q < 30$ for $d \approx 10\mu$.

already irregularly distributed in the packed powder, or whether the separation occurred at the moment of dispersion could, of course, never be detected by our method. But, since the powder as a whole was neutral, and since any charge is effective only when separated from its opposite, this question is of little interest. We could, also, assume that the charge distribution reached some kind of an equilibrium after comparatively little friction, so that any further impacts and additional friction would not change the resultant charge distribution noticeably. The difference between these two interpretations can be summarized in the following way: In the first cases, the main charge separation is supposed to occur at the first mechanical separation of the particles; whereas in the latter, it is regarded as a result of several collisions and possibly friction of very short duration. However, as later interpretation shows, the evidence is all in favor of a charge distribution on separation of contacting surfaces. There is, of course, a small chance that the strong variation in turbulence, contrary to expectation, did not effectively change the amount of friction and the number and violence of the impacts between particles. In that case, it would be expected that no difference in the observed particle charges would be found. At any rate, the electrostatic cloud characteristic shown in Fig. 8 was surprisingly independent of the conditions of dispersion.

Changing the moisture content of the surrounding atmosphere in the hope of varying the water layers on the dust particles, and thus the charging, produced no material change in the type of distribution shown in Fig. 8. But, as already pointed out, the water content of a fine non-hygroscopic powder was known to be remarkably unaffected by the relative humidity of the surrounding medium, so that any significant variation in charging due to moisture was hardly to be expected.

A short summary of the most important experimental results is given in Table II.

There it is seen that still more surprising than the invariance of the charges generated in homogeneous systems, was the fact that many so-called inhomogeneous systems showed nearly the same symmetric charge as in Fig. 8. When quartz, talc, or starch was blown out of a pyrex duster, the charges on the particles balanced just as before* and when sulphur was blown out of a quartz, Pyrex or nickel tube, the predominance of one sign of charge was only very slight.

Again, two different explanations are possible. First of all, there was little doubt that only a small fraction of particles actually hit the solid walls in the process of generation, most of them being surrounded by a layer of air all the time, and thus being cushioned by the air and failing to strike the walls. Larger particles had a better chance of impact than smaller particles because of their inertia. Thus, most of the charges were generated by dispersal into the cloud or by collisions between particles. Those particles that hit the walls either did not interact sufficiently with the wall lattice to cause any appreciable charge transfer, or else the differences in surface characteristics were not large enough on the average to disturb the initial symmetry perceptibly. Strangely, the sulphur powder did not show a very large asymmetry, even when forced through the nickel duster with as many as eight baffles. In that case, certainly, most of the larger particles must have hit the sides. We have not been able to account adequately for this fact, except, possibly, by noting that the observed samples could not be made large enough to yield reliable statistics. (The sulphur was not easily dispersible.)

The talc powder, when blown through the nickel chamber, did not show the slightest asymmetry in its

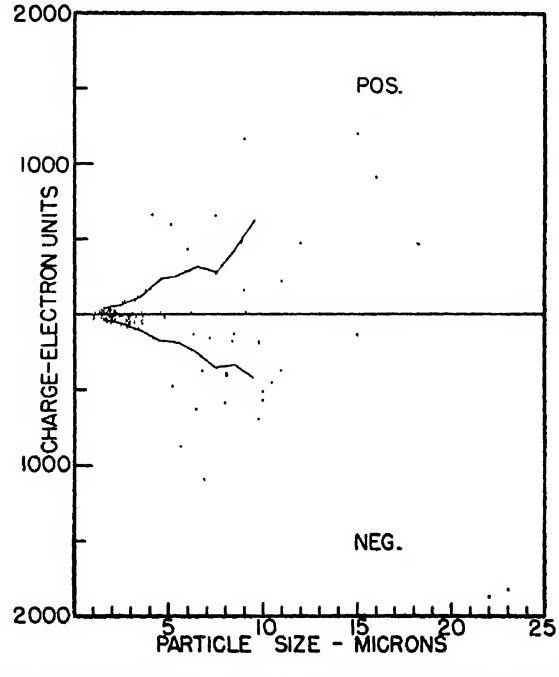


FIG. 8. Charge *vs.* size representation of a cloud of quartz powder blown out of a quartz cup. The average charges per particle are drawn in; the spread is seen to be large.

* A similar result was obtained by Hansen with $CaCO_3$ powder in a SiO_2 duster.

charge distribution. Here, however, the explanation was obvious. Opening the nickel duster after the experiment revealed that the inside was completely lined with a heavy layer of talc powder. This layer could not be blown off but had to be wiped away. Apparently, all talc particles hitting the metal sides stayed there, possibly attracted by their image charges, until all metal surfaces were covered, thus effectively reducing the conditions to those of the homogeneous cases.

The only fairly pronounced asymmetry in charge distribution involving solely non-conductors was obtained when a fine quartz powder was blown out of a sulphur lined duster. The negative sign predominated distinctly, which was in qualitative agreement with the fact that in the sulphur powder dispersed out of a quartz duster the positive charge was noticeably larger.

It was noted, at the same time, that the asymmetry was stronger in the larger particles, being very slight for the smaller sizes. All other strongly asymmetric cases contained metallic elements. When insulating powders were used with metal or vice versa, this insulating dust invariably acquired a predominantly negative charge, leaving the metal with the corresponding positive balance. The largest deviation from the symmetrical distribution occurred in a quartz cloud blown out of the platinum dispersing chamber with eight baffles. A plot showing every other particle is given in Fig. 9. It is seen that the negative charges predominated both in number and magnitude. Still a large number of positive particles, especially of small diameter, were present. None of the larger particles carried a positive charge. Evidently the quartz in contact with platinum was likely to acquire a surplus of electrons.

All large particles or glomerates hit the metal surfaces at least once, whereas many of the smaller ones never touched the walls, being carried along by the air stream well imbedded in the gas. Many of the small particles that succeeded in colliding with the sides of the chamber stuck there, forming a thin coating. A crude measurement of the positive charge imparted to the duster showed this to be of the same order of magnitude as the total net charge estimated to be present in the cloud. This can be considered as evidence that the observed negative surplus charge was not balanced by invisible positive particles left behind in the settling column. If desired, this conclusion could, of course, very easily be verified by employing any of the many well-known space charge detecting devices.

In agreement with these findings, a nickel powder blown out of the quartz dusting tube carried a predominantly positive charge with a spread otherwise similar to that shown in Fig. 9. Either the nickel particles, unlike those of sulphur, had a large chance of touching the sides of the tube, possibly because of their larger inertia for equal volumes (higher density), or the charge transfer between nickel and quartz is easier than that between sulphur and nickel. The

TABLE II.

Symmetric					Asymmetric				
Powder	Container	Sign	Deviation[c]	St. dev.[c]	Powder	Container	Sign	Deviation[c]	St. dev.[c]
SiO_2	SiO_2	−	0.035	0.05	SiO_2	S	−	0.21	0.11
S	S	−	0.036	0.04					
Starch	Pyrex	−	0.020	0.12	Ni	SiO_2	+	0.21	0.07
Talc	Pyrex	+	0.015	0.08	Sb	Pt	−	0.60	0.21
SiO_2	Pyrex	−	0.10	0.12	Ni	Pt	+	0.34	0.10
Talc	Ni[a]	−	0.06	0.11	Ni	Ni	+	0.32	0.09
					Starch	Ni	−	0.33	0.11
Slightly asymmetric					SiO_2	Ni	−	0.39	0.07
S	SiO_2	+	0.13	0.11	SiO_2	Pt	−	0.50	0.11
S	Pyrex	−	0.13	0.11					
Sb	Ni	−	0.19	0.16					

[a] Ni duster found coated with a layer of talc powder.
[b] Sulphur powder could not be well dispersed.
[c] The deviations from symmetry are expressed by the ratio $x = \frac{N_+ - N_-}{N_+ + N_-}$.

nickel powder, when dispersed out of the nickel duster, however, did not behave like the other homogeneous systems, very probably because the surfaces of the reduced Ni particles differed appreciably in their properties and constitution from the nickel plate walls. This was despite the fact that every common precaution possible had been taken to insure uniformity and zero contact potential.

As shown in Table II, an observed distribution is considered to be symmetric when the deviation from the exactly balanced distribution is less than the statistical standard deviation of the particular experiment. Similarly, if the deviation is only a little larger than the standard deviation, the distribution is called slightly asymmetric. If it is much larger, the result is regarded as asymmetric. In this way, the statistical factor, which is unavoidably connected with any measurement on random samples, is at least partially taken into account. This procedure of classifying the results merely according to the relative numbers of positive and negative particles instead of according to the magnitude of these charges is justified, since all experiments show a strong correlation between the relative numbers and the average charges of the particles of opposite sign.

Finally, in order to demonstrate most convincingly that much of the charge separation takes place during the very first phase of the dispersion, the fine quartz powder was mixed with the easily lumping sulphur and blown very, very gently out of a quartz cup by a slow vertical stream of air. In this case, the large particles seemed to be fairly randomly charged with only a slight excess of the negative sign; whereas the small particles which arrived later proved to be overwhelmingly negative. These latter were certainly mostly quartz; the former must have consisted partly of the large sulphur glomerates. Since both the contact between particles and dissimilar walls and the turbulence and friction between particles was presumably negligibly small, the charge separation must have become effective at the very first breaking of the close contacts between the dissimilar surfaces of the quartz and sulphur particles. It should be noted that in

cases like these, all elements of a static machine separating or segregating charges which could ultimately lead to sparks are present. The dissimilarity of large and small particles results in an asymmetry of charge by small particles having one sign larger than the other. Mechanical separation in the gravitational field or by air currents then leads to segregation of charges and local accumulation.

CONCLUSIONS

There appears to be no doubt that the act of separating a mass of particles of solid in contact with each other into a cloud of assorted sizes of particles, yields in all cases a cloud which as a whole is neutral, but in which practically all the particles are charged. In a homogeneous substance, where contact with heterogeneous surfaces are at a minimum, the numbers of positive and negative particles in each size group from about 0.5 micron to 30 microns in diameter are about equal. The average charge increases somewhat more slowly than proportional to the surface of the particles. Apparently the character and forces involved in the separation and the possibility of subsequent friction in turbulent clouds with violent blowing produce no notable change in the electrification, nor does the charge increase with increase in contact or the energy involved in the process. In the separation of an originally mixed dust of different average size for the two components, the study of the cloud indicates that the electrification occurs primarily at the instant of separation of the two components. If homogeneous clouds are blown so that they make contact with solid surfaces of a different sort in sufficient measure to affect a notable fraction of the particles, the separation of the heterogeneous system of dust and solid surface may impose an asymmetry on the equality of positive and negative particles from the dispersed cloud particles that did not contact the solid surface. Results are reproducible in any given system. However, in a surprisingly large number of impacts between dissimilar surface and dust, no great influence was observed. In general, the asymmetry is exhibited more strongly by the larger particles, which, because of their mass, were more likely to strike the solid surface. Very small particles of the order of one or two microns appear to be cushioned by the gas and do not make contact with the surface in some cases. In occasional cases, e.g., talc on metal, care must be taken in interpreting the results, since the dust may be so highly charged and insulating that it is held to metal surface and most of the cloud makes contact with the dust coating only, showing the charge distribution of a homogeneous substance. The direction of charge transfer between heterogeneous surfaces is unpredictable and depends on the substances involved. With some insulators, the character of the surface, whether crystalline or amorphous, appears to cause no asymmetry or charging. Thus, amorphous quartz and crystalline quartz showed no asymmetry. Likewise, Pyrex glass, a borosilicate glass, and quartz showed no asymmetric transfer. The most universal asymmetry observed was that of metal and insulators, where the insulator universally received a negative charge. In the case of a Ni powder reduced in H_2 against Ni or Pt sheet strong asymmetry was observed with the reduced Ni powder losing electrons to Pt and to the sheet Ni. This is to be expected in the case of Ni and Pt, for Ni has a low work, function relative to Pt. In some cases where asymmetry, owing to contact of dust and surface, predominated, the charge on the duster roughly equalled the total net charge of the opposite sign on the cloud. These results, as a whole, and especially those with metals, not only through their consistency and reproducibility, but through the signs of electrification obtained, give considerable confidence in the use of this method and the results it yields. In fact, it now appears that the method is ready to be applied to systems of exceptional purity and in carefully controlled atmospheres, in order to derive some hints as to the true character of the processes of charge exchange in non-homogeneous systems as related to the solid state.

Insofar as these studies go, they were done on reasonably clean surfaces and a fairly representative spread of substances to cover the field. The ambient gas was fairly dry and filtered air, subjected to the natural ionization by radioactivity in the surroundings and cosmic rays in the settling tower. Probably, in this connection, one of the most striking discoveries in the

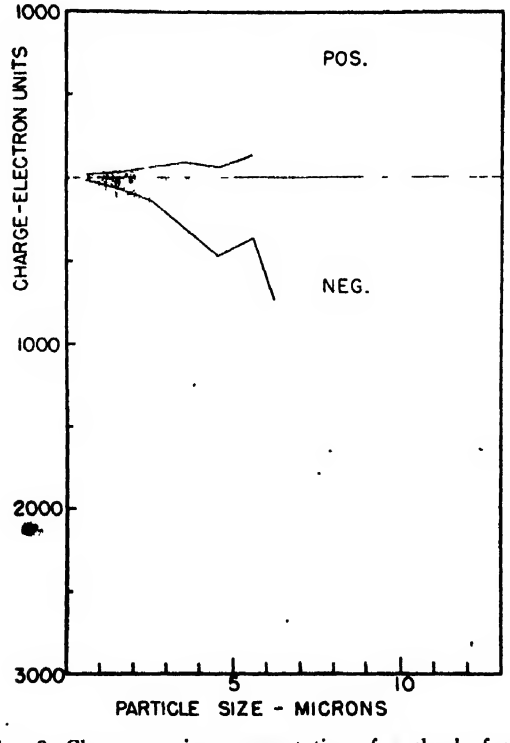


FIG. 9. Charge *vs.* size representation of a cloud of quartz powder dispersed out of a platinum duster. The average charges per particle are drawn in; note the large spread.

work with the normally dry dusts used, was the minor role played by the water vapor and moisture. Since trouble from this source was expected, considerable time was devoted to its study. Actual studies, especially on the quartz, indicated that if heated above 500°C, it lost all superficial moisture: On cooling to room temperature, or even to 120°C, it rapidly absorbed a small amount of water. By weight, it was equivalent to no more than enough to cover the average particle if spherical with, at most, a layer two molecules deep. From the electron microscope pictures of the particles of quartz, the surface is really so much greater than assumed that the layer could not be more than one molecule thick, and then the H_2O molecules may only occupy preferential spots. Changing relative humidity up to 90 percent at 20°C did not alter the layers at all. Above this, the surface became wet and the dust stuck together. Such dust did not disperse. From this it must be concluded that the water played a minor role in the electrification. Certainly, there was not enough water to give aqueous films, so that the classically conceived electrolytic ion exchange could account for the electrification.

It is probably too early to try to evolve a picture of the processes at work. However, the universal symmetry of charging on separation of homogeneous particles after contact and the asymmetrical charging of some of the heterogeneous systems, together with observations in the electron microscope, permit us to formulate a tentative theory of what happens. In the case of the homogeneous system, it is clear that the major portion of the charging comes from the separation of particles initially in contact. The very great sub-microscopic roughness of the surfaces which was observed, together with no evidence that this roughness constitutes a fine sub-microscopic dust which charges independently of the microscopic particles in the cloud, leads one to expect relatively strong pressures in contact points between the major aggregates microscopically observed. These pressures could *eliminate gaseous or surface films* and thus constitute intimate contact *between small areas of the homogeneous substance.* The separations on dispersal may take place along the actual region of contact, or they may take place by the actual fracturing of the small contacting points, depending on how intimately the pressure has welded the contacting points of the two larger pieces together. In any case, separation and dispersal into a cloud involves producing some sort of rupture between a large array of minute points. Doubtless, these points are of various sizes. If now one makes the assumption that for a contact area of some minimum size, i.e., radius σ involving, say, some 10^3 or 10^4 atoms of the crystal lattice, there is a good chance that crystal imperfections will leave one extra electron or ion of negative charge lying near the plane of fracture on one side of that plane; then that particular point of contact on breakage will contribute a negative charge to its large particle and the equivalent positive charge to the other particle. Such chances of separation are easily possible, due to imperfections in the lattice in which a negative ion or electron is out of place and is geometrically separated from the equivalent positive charge relative to a cleavage plane. Cleavage along such a plane will then cause a charge separation.

On this basis, statistical analysis enables us to compute the probability that a single large particle will have a certain excess charge, positive or negative. The larger the surfaces of the large particles in contact, the more points are ruptured and from pure probability considerations using the concept of a minimum area per unit charge imperfection, the statistics and relations may be derived.

If the powder contains a continuous spectrum of sizes, the contact area of a certain particle with its neighbors will be on the average proportional to its surface.

Let us define a *unit contact* as a region of extremely high probability (certainty) of transfer of one electron on particle separation. Thus, at each unit contact one particle is left with an excess and the other with a deficit of one electron of charge. Let us further assume that the number of such unit contacts is large on all particles to be considered and that the direction of charge transfer is completely random in any homogeneous powder.

Let the number of unit contacts of a certain particle be N. Thus, $\bar{N} \propto S$ (S=surface area of particle). If the number of electrons gained is K ($K<N$), then the number of electrons lost at the other contacts is $N-K$ and the net charge finally will be $n=(N-K)-K=N-2K$ elementary charges. The probability of such an event is given by

$$P_K{}^N=\frac{N!}{2^N(N-K)!K!}=\frac{N!}{2^N}\cdot\frac{1}{(N/2+n/2)!(N/2-n/2)!}.$$

(This assumes exactly equal chances of electron loss and electron gain.) Since N is large, we can use Stirling's formula and obtain

$$P_n{}^N\cong\left(\frac{2}{\pi N}\right)^{\frac{1}{2}}\frac{1}{[1+(n/N)]^{(N+n+1)/2}[1-(n/N)]^{(N-n+1)/2}}.$$

If N is large, such a binomial distribution curve can be very adequately approximated by $P(n)=(2/\pi N)^{\frac{1}{2}}\times\exp(-n^2/2N)$. The plot shown in Fig. 8, indeed, suggests such a Gaussian distribution of charge for constant particle size. The average charge will then depend in a simple fashion on the size as can be seen from the following discussion.

$$\bar{n}=(2/\pi N)^{\frac{1}{2}}\int_0^\infty n\exp(-n^2/2N)dn=(2N/\pi)^{\frac{1}{2}}.$$

Let $N=(\pi/2)\alpha^2 S$. Then $\bar{n}=\alpha(S)^{\frac{1}{2}}$. Here α is a constant factor, assuming that the statistical relation between N and S does not disturb the average. Thus, this theory

leads to the conclusion that the average charge per particle of one sign should be proportional to the square root of the particle surface.

For spherical particles $\bar{n}=Bd$ with d the diameter. Since, for the cases investigated, the particle shapes were very far from spherical, no such simple relation can be expected. Inasmuch as the irregularities always tend to increase with size, owing to the formation of aggregates, we should expect $\bar{n}$ to vary somewhat more rapidly than linearly with d which is exactly what is observed. This simple theory of random charge transfer at points of close contact then leads to conclusions in good agreement with observations. It must be noted that while the Gaussian distribution for the symmetrical cloud as deduced has a maximum at zero charge, the observations on, say, 100 particles in a given size range with some 1000 charges would not show the large number of uncharged particles of zero charge on purely satistical grounds. If this picture is accepted, the values listed in Table I will at once yield an estimate of the number of unit contacts on particles of a certain size. Thus, for the particles of diameters between 2 and 3 micron, the data give N as approximately $N=\pi/2\bar{n}^2 \cong 4000$. For those particles between 0.5 and 1 micron, the value of N is approximately $N\cong 400$.* Since, on the other hand, the surface area of the particles is known, at least very roughly, the maximum average size of a unit contact, as defined above, can be estimated. It is found from our data that for the quartz used by us, at least one electron is transferred at each area of contact of 10^{-10} cm^2. Similar magnitudes occur for all the particles observed. The differences existing are hidden in the uncertain knowledge of the surface areas involved. These values are maximum estimates. Actually, the average inhomogeneity at surfaces can well be confined to smaller areas, since only a small fraction of the total surface will be in actual contact. If the contacts are confined to 1/100 of the total surface, the average area of a unit contact will be 10^{-12} cm^2, which corresponds to about one imperfection for every 10^4 atoms, a figure that is more likely than one in 10^6.

In the case of heterogeneous surfaces, differences in structure, as well as in surface layers, will result, on the average, in charge transfers at contact points which can well be asymmetrical. This would be particularly likely where metals having free electrons can give these off to the ionic and other lattices having electro-negative elements. Even then, many of the contacts can be separated by fracture in the homogeneous substance leading to more homogeneous charging. Therefore, it will be expected that even in the most heterogeneous systems, the asymmetrical charging will never be complete. This is especially true, of course, where friction with dissimilar surfaces, as in the duster with many baffles, does not occur. It is probable, in view of what has been said, that it will be found that in heterogeneous systems the gas films and surface conditions play an important role, whereas they do not do so in the more homogeneous systems where possible pressure fusion and rupture of the minute points is more likely.

I am indebted to Professor L. B. Loeb, at whose suggestion this problem was undertaken and under whose supervision it was carried out, for his many valuable suggestions and assistance in interpretation, which were a constant source of inspiration. Also, I wish to thank Mr. R. J. Wijsman for his stimulation and verification of the statistical theory of mutual charging here presented.

* The observed average charge per particle was not corrected for neutralization by air ions in this case. Small particles, which were often photographed at a considerable time after their separation from the powder, would be expected to have lost some of their initial charge. Thus, the true value of N can be expected to be somewhat larger.

Charge Distribution in Coarse Aerosols as a Function of Time*

W. B. KUNKEL
Department of Physics, University of California, Berkeley, California
(Received April 3, 1950)

The interaction of small particles suspended in air with the ions normally produced in the air is discussed. The neutralization of highly charged dust particles, if suspended in air, is treated in detail. The result is found to be in qualitative agreement with experiment. The special case of final equilibrium is investigated. It is found that multiply charged particles should be present in appreciable number in coarse aerosols at all times. This is confirmed by experiments both for initially charged quartz dust and for initially neutral ammonium chloride smoke. The general nature of an equilibrium charge distribution is presented. Charges up to ten electron units are found to be not uncommon.

INTRODUCTION

IT has long been known that microscopic and submicroscopic particles and droplets suspended in air frequently carry electric charges.[1] Among others, Patterson in his extensive studies of aerosols[2] noted the fact and developed a theory involving the coagulation taking place in fresh aerosols of high density. Experimentally, Patterson and his collaborators found that the percentage of charged particles in originally uncharged clouds increased with time, the numbers of positives and negatives remaining equal throughout. Originally charged sols would remain charged. The magnitude of the charges were, however, not determined.

The processes by which these charges are acquired and/or altered can be divided into three groups. First, the violent creation of an aerosol, be it by a chemical reaction, by an electric arc, by the mechanical action of an explosion, or merely the dispersion through a blast of air, will in general leave the particles charged.[3] Such charging may be by triboelectric actions, by spray electrification, or possibly by direct ionization or thermionic emission due to strong local heating. These actions will determine the initial charges of the particles. Secondly, there may then be subsequent collisions between particles involving frictional charging, charge transfer and agglomeration. Such contacts caused by Brownian motion and turbulence in the medium can appreciably alter the initial distribution of charges in time. Finally, there will always be direct exchanges with the ions present in the ambient air, thus effecting a change of charge on the particles even when there is no contact between particles.

This study concerns itself with the effect of ionic charge exchanges of relatively large particles which are practically unaffected by Brownian motion, and it is limited to cloud densities small enough to render the probability of mutual impact of the particles negligible.

* This work was supported by an ONR Contract.
[1] H. Freundlich, "Kapillarchemie," Akad. Verl. Ges. (1922).
[2] Patterson, Proc. Roy. Soc. A124, 523 (1929). Also Phil. Mag. 12, 1175 (1931).
[3] H. Sachsse, Ann. d. Physik 14, 396 (1932).

CHANGE OF CHARGE OF SUSPENDED PARTICLES

A small particle suspended in air is continually bombarded by gas molecules. Among the molecules there are a certain number of gaseous ions of both signs produced continually by various processes. These ions will have a finite chance of attaching to the particle or at least of being neutralized on impact, thus transferring their charge to the particle. If the particle was initially charged, this change of charge by one unit may represent either an increase or a decrease. The probability of such an event will depend primarily on the initial charge of the particle, on the ion density in the air, on the size, shape and material of the particle, on the temperature, and, if present, on any relative directed motion of the ions; e.g., on the average drift velocity of the ions, if there is an electrical field acting as in Millikan's oil drop study.

It is also possible for a particle to change its charge due to an ionizing process directly affecting the particle itself; e.g., caused by some radiation striking it, and ejecting one or more electrons. The probability of such an event will depend on the size and material of the particle, the total surface of the particles relative to the absorbing cross section of the gas molecules, and on the character, energy and amount of ionizing radiation. Owing to the values of the relative numbers of molecules to those of the particles, the direct ionizing effects for many types of radiation will be much smaller than the charging because of the ionization of the air. It is clear that the problem presented in charging of dusts by natural ions in the air as indicated above involves an equilibrium between the ions of both signs present in relatively small equal amounts, a problem apparently not solved before. On the other hand, the accumulation of charge on particles suspended in regions of relatively high ion densities of one sign as affecting electrical precipitation has been studied by many authors, notably by Arendt and Kallmann.[4]

In order to study the capture of ions in the air by particles such as are under discussion, we proceed as follows. Let us call $\varphi_1(q)$ the probability that a particle of charge q and of a certain size, shape and substance

[4] Arendt and Kallmann, Zeits. f. Phys. 35, 421 (1925).

suspended in a medium of a certain ion density and temperature gains an electron unit of charge in unit time. And let $\varphi_2(q)$ be the probability that an identical particle under identical conditions loses an electron unit of charge per unit time. Furthermore, let N_q be the number of such particles suspended per cubic centimeter. Then the change in the density of such particles in still air due to charge transfers, neglecting the settling motion, becomes:

$$\Delta N_q=[N_{q-1}\varphi_1(q-1)-N_q\varphi_1(q) -N_q\varphi_2(q)+N_{q+1}\varphi_2(q+1)]\Delta t. \quad (1)$$

If particles of various sizes are present in the volume element similar equations hold for each infinitesimal size range separately. Since agglomeration and evaporation or condensation are considered to be negligible, there is no change of particle size and these equations are independent of each other. Yet, inasmuch as φ_1 and φ_2 are generally unknown, Eq. (1) is of no great importance. Only in very special cases will certain simplifications enable us to make predictions.

NEUTRALIZATION OF A CHARGED SMOKE

If the individual particles of a cloud of smoke or fog are highly charged, and if the densities of positive and negative ions are of the same order of magnitude, φ_1 becomes negligible compared to φ_2 which will consist almost exclusively of the probability of neutralization by air ions. Moreover, since $q\gg1$, we can treat the changes of charge as infinitesimals. Thus we can write

$$\frac{\Delta N_q}{\Delta t}\rightarrow\frac{\partial N_q}{\partial t}dq=-N_q\varphi_2+\left(N_q+\frac{\partial N_q}{\partial q}dq\right)\left(\varphi_2+\frac{\partial\varphi_2}{\partial q}dq\right)$$

or

$$\frac{\partial N_q}{\partial t}=N_q\frac{\partial\varphi_2}{\partial q}+\varphi_2\frac{\partial N_q}{\partial q}=\frac{\partial N_q\varphi_2}{\partial q}, \quad (2)$$

where $\varphi_2(q)=-dq/dt$ is the average rate of loss of charge of a particle of charge q.

Equation (2) is of practical value only if $\varphi_2(q)$ is a relatively simple expression. To obtain a first approximation, we can assume that, on the average, ions, due to their heat motion, entering a sphere of radius r around a particle of charge q will be trapped, r being given by the relation

$$E_{\text{pot}}+E_{\text{kin}}=\frac{qq'e^2}{r}+\frac{3}{2}kT=0. \quad (3)$$

Here e is the electron charge, q the number of electron charges on the particle, q' the multiplicity of the ion charge, k the Boltzmann constant, and T the absolute temperature. The quantity r is positive only if q and q' are of opposite sign. This relation assumes a Maxwellian velocity distribution of the ions. Then the number of neutralizing events per second will be given by the well-known relation

$$\varphi_2=-dq/dt=n\pi r^2\bar{c}=g(t)q^2 \quad (4)$$

again to a first approximation. Here n is the density, $\bar{c}$ the average velocity of the ions, and $g(t)$ is a factor independent of q derived from Eqs. (3) and (4)

$$g(t)=\frac{4\pi\bar{c}e^4q'^2}{9k^2T^2}n(t)=\beta n(t). \quad (5)$$

It should be noted that in this case φ_2 is independent of the size of the particle as long as its radius $\rho<r$, so that the treatment will hold even if a cloud contains particles of various sizes as long as the cloud density does not change due to the difference in settling rates.

Substituting (4) and (5) into (2) we obtain

$$\frac{\partial N_q}{\partial t}=\beta n\frac{\partial}{\partial q}(q^2N_q). \quad (6)$$

This means that q^2N_q is independent of time if $q(t)$ is regarded as the charge on a particular particle changing according to (4):

$$q=\frac{q_0}{1+\beta q_0\int_0^t ndt}. \quad (7)$$

Thus

$$N_q(q,t)=(q_0^2/q^2)N_q(q_0,0), \quad (8)$$

where $N_q(q,0)$ is the initial charge distribution in the cloud and q_0 the initial charge of a particle of charge q at time t:

$$q_0=\frac{q}{1-\beta q\int_0^t ndt}. \quad (9)$$

The problem is therefore reduced to that of finding $n(t)$, the ion density in the region of the cloud to be considered. Unfortunately this cannot be solved rigorously since $n(t)$ is determined by the recombination equation:

$$dn/dt=\gamma-\alpha n^2-\beta n\int_0^\infty q^2N_q dq, \quad (10)$$

assuming, for the time being, equal numbers of positive and negative ions present. Again γ is the rate of formation of ion pairs per cubic centimeter and α the ordinary coefficient of recombination. α is very small so that in the absence of charged dust, n becomes fairly large until equilibrium is reached, but if charged particles are suspended in the air, the value of n is correspondingly depressed. Only certain extreme cases can be treated in more detail. If the aerosol is very dilute, the last term in Eq. (10) is very small compared to αn^2, and n will be practically the same as in dust-free air. It can there-

fore be regarded as constant in time if equilibrium between a constant ion production and recombination is established. This case is very simple since the integral in Eq. (9) is trivial.

On the other hand, if q is large and N_q not very small, the last term in Eq. (10) will easily predominate so that αn^2 effectively can be neglected. This means that practically all ions formed will attach to the smoke particles. It also presupposes that the suspended particles carry both positive and negative charges in about equal amounts, a case which seems to be frequently realized.[3] Moreover, if we make use of the fact that in this simplified case the rate of change of ion concentration is small compared to the ion production, we can reduce (10) to

$$0 \cong \gamma - \beta n \int_0^\infty q^2 N_q dq, \tag{11}$$

whence

$$\varphi_2 \cong \frac{\gamma q^2}{\int_0^\infty q^2 N_q dq} = -\frac{dq}{dt}. \tag{12}$$

The integral in Eq. (12) is a function of time and since the integrand is in general not known in detail, further

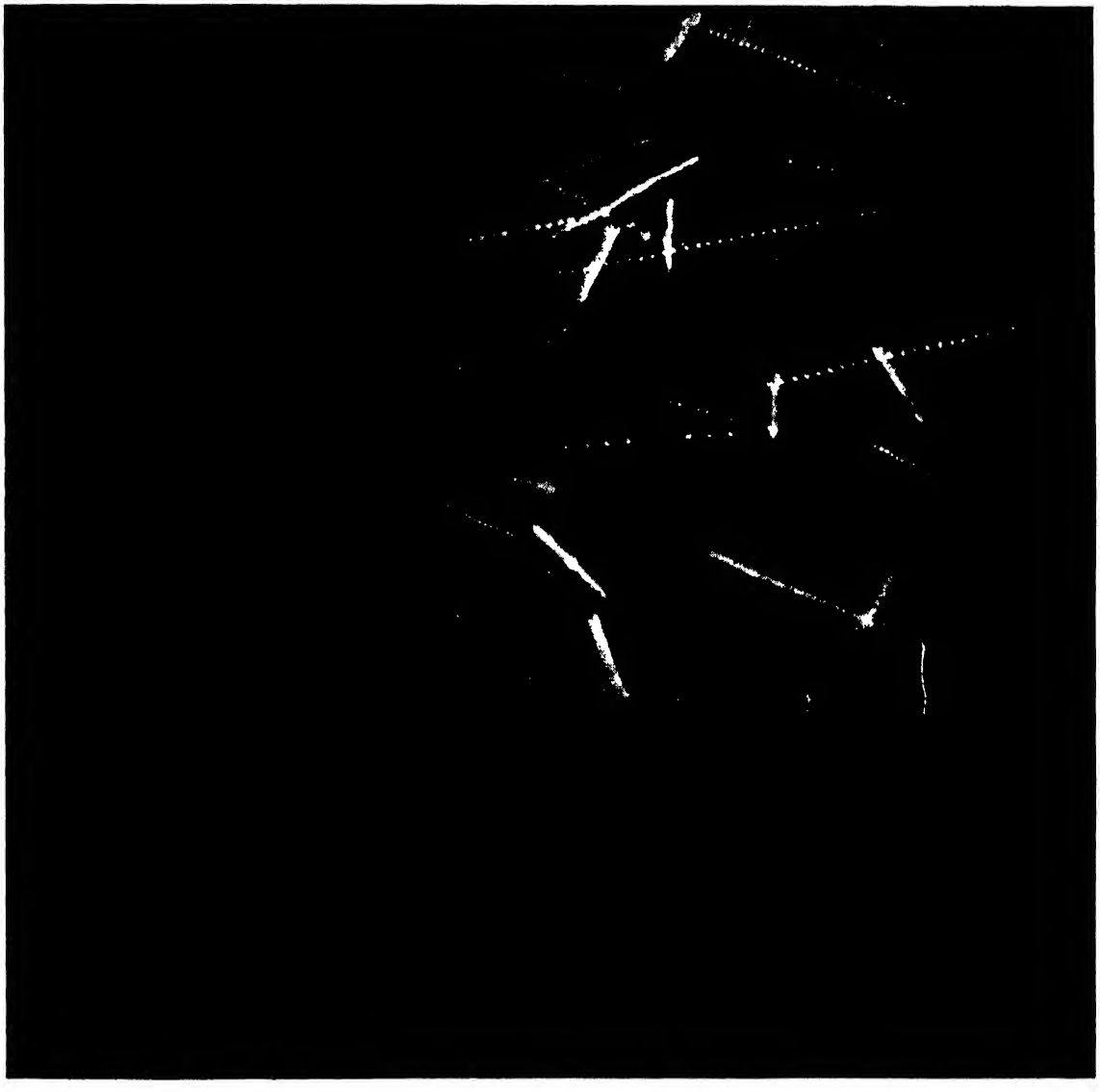


FIG. 1. Traces of quartz dust, 20 hours old, settling in air in a horizontal field of 560 volts cm^{-1}. Dark field illumination, 450 flashed $min.^{-1}$. Total exposure about 4 sec. Neutral particles settle vertically. Singly and doubly charged particles of both signs are easily recognized. For higher multiplicities the traces have to be carefully measured, for possible convection correction has to be made. Note the distinct effect of Brownian motion or differential heating on these irregularly shaped particles.

generalized development is impossible. However, by defining a root mean square charge

$$q^*(t)=\left[1/N\int_0^\infty q^2N_q dq\right]^{\frac{1}{2}} \tag{13}$$

and substituting this into (12), inspection shows that it is justifiable to write for the characteristic charge of the cloud $(dq^*/dt)=(-\gamma q^*/Nq^{*2})=-\gamma/N$ which yields

$$q^*=q_0^*-\gamma/Nt. \tag{14}$$

In other words, the r.m.s. charge in a symmetrically charged smoke cloud decreases linearly with time as long as the approximations introduced are justified. This is a rather remarkable result since the average charge of the smoke particles,

$$\bar{q}=(1/N)\int_0^\infty qN_q dq=Q/N \tag{15}$$

by our own definition decreases at the same rate:

$$d\bar{q}/dt=1/N(dQ/dt)=-\gamma/N,$$

where Q is the total charge of one sign per unit volume of the cloud.

Combining Eqs. (12)–(14), the most probable rate of change of charge of each particle is expressed as a function of its charge q and of the time, containing the initial r.m.s. charge q_0^* of the cloud element as a parameter:

$$\frac{dq}{dt}=\frac{-\gamma q^2}{N[q_0^*-(\gamma/N)t]^2}. \tag{16}$$

This is readily integrated, resulting in an expression analogous to (7), to wit:

$$q=\frac{q_0}{1+(\gamma q_0 t/Nq_0^{*2}-\gamma q_0^* t)}, \tag{17}$$

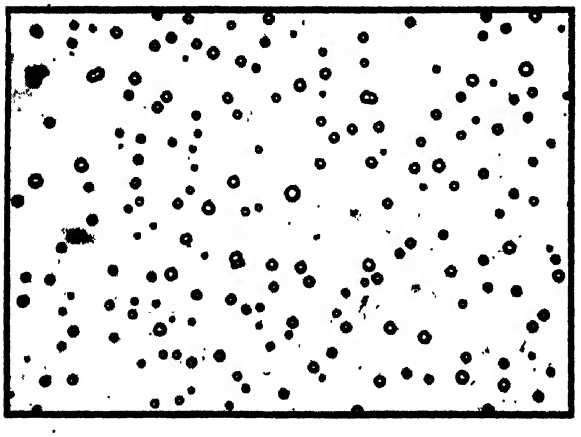


FIG. 2. Ammonium chloride smoke particles collected on microscopic cover glass at bottom of the dusting chamber. Sizes range from 0.7 to 3 microns in diameter.

or

$$q_0=\frac{q}{1-(\gamma qt/Nq_0^{*2}-\gamma q_0 t)}, \tag{18}$$

which has to be substituted in Eq. (8).

Unfortunately we have not been able to verify these simple relations quantitatively since, in our experimental arrangement,[5] it was not possible to keep a smoke homogeneously suspended for a sufficiently long period. Qualitatively, however, these ideas appear to be in good agreement with observations discussed elsewhere.[6]

EQUILIBRIUM DISTRIBUTION

The approximations introduced above are justified only if the charges are fairly large, say, with an average of 50 or more electrons per particle, and if the aerosol is not too dense. As the charges of the particles decrease, the ion-ion recombination gains in importance so that Eq. (11) can no longer be used and the second treatment breaks down. All we can say is that although the ion concentration will gradually increase, the rate of neutralization of the smoke will decrease more and more until φ_2 becomes comparable to φ_1 so that the latter no longer can be neglected. In other words we have to return to Eq. (1). Moreover we shall no longer be able to make use of differentials so that a rigorous treatment would be extremely complicated. Again, only the very extreme case of negligible change of charge distribution can be sufficiently simplified to permit a few quantitative conclusions. In this case the conditions of balance tell us that not only $\Delta N_q=0$ but that the terms on the right hand side of Eq. (1) have to cancel in pairs, i.e.,

$$N_{q-1}\varphi_1(q-1)-N_q\varphi_2(q)=0, \tag{19}$$

if the equilibrium is to be stable. Thus the equilibrium distribution can be expressed in the form

$$N_q=N_0\prod_{l=0}^{l=q}\frac{\varphi_1(l-1)}{\varphi_2(l)}, \tag{20}$$

if we define $\varphi_1(-1)=\varphi_2(0)=1$. Since φ_1 and φ_2 depend on the size of the particles, each size range will have a different distribution unless the size dependence cancels out in the expression. It is clear that $\varphi_1(q)$, the probability of charge addition, will decrease as q increases. The question arises whether $\varphi_1(2)$ for most smokes is not already negligible so that practically no particles with more than two elementary charges will be present in the cloud near equilibrium. Unfortunately many factors contributing to φ_1, such as the effect of the material and the shape, are largely unknown so that quantitative estimates cannot be expected to be reliable. We can assume, however, that the charges on a spherical

[5] W. B. Kunkel and J. W. Hansen, Rev. Sci. Inst. 21, 308 (1950).

[6] W. B. Kunkel, J. App. Phys. 21, 820 (1950).

particle arrange themselves on its surface at maximum equal distances (conducting droplets) and that all ions actually colliding with the surface will impart their charge to it. This very crude first approximation leads to the conclusion that in a smoke consisting of particles of about one micron diameter, multiple charges as high as five or even ten electrons are still quite likely.

This rather unexpected result is in qualitative agreement with the following experimental evidence. A small amount of fine quartz powder was dispersed by a short blast of air. The particles of the dust cloud thus generated were found to be quite highly charged, positive and negative charges being present in about equal quantity. The charges and sizes of the particles were determined by photographing their trajectories as they settled in a horizontal electric field with dark field illumination. The time scale was provided by using intermittent light of known flash frequency. This method has been described in some detail elsewhere.[5] After about one hour, all particles larger than one micron in diameter had settled out. Most of the small particles still present carried charges of between 10 and 20 electron units of either sign, the total range being from zero to about 100 electron units. A minute convection kept most of these particles suspended for 24 hours, after which time the charges were measured again. It was found that the average charge had decreased to about 3 electron units, charges of higher multiplicity still being present in large number. An example of the pictures of the falling particles obtained is shown in Fig. 1. There appeared to be no significant change in this distribution during the following few hours for which the suspension could be maintained.

In order to confirm the theoretical considerations above, it was next demonstrated that the charges observed were not merely residues of the initial contact electricity but represented a true equilibrium condition. A cloud of ammonium chloride smoke was generated by mixing ammonia and hydrochloric acid vapors in room air of 40 percent relative humidity at 20°C. The particles formed under suitable conditions were fairly uniform in size and quite round as seen in Fig. 2. Apparently aggregation between large particles did not take place, the growth consisting entirely of addition of very small particles or of single gas molecules of NH_4Cl to solutions of NH_4Cl in water. The surfaces appeared to be quite smooth and spherical under the electron microscope. A few minutes after complete formation of the cloud, analysis showed these particles to be mostly neutral, a certain fraction being singly charged. Three hours later the percentage of charged particles had markedly increased, multiple charges being already present in appreciable number. Again a part of the smoke was kept suspended by a slight convection in the chamber. The distribution found on the next day is shown in Fig. 3. The particles were grouped according to their charge irrespective of their size. Actually the average charge per particle increased somewhat with

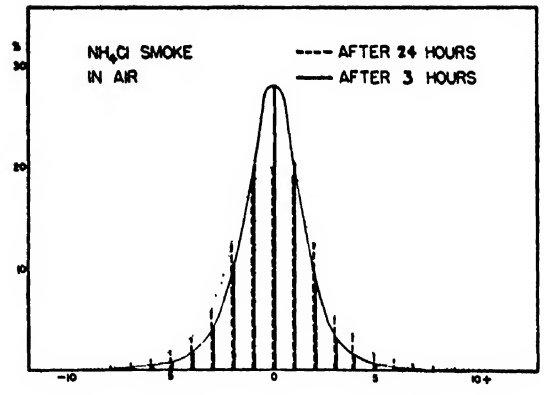


FIG. 3. Charge distribution in ammonium chloride smoke made by mixing the gaseous components. The curves are based on totals of about 1000 particles each.

the size of the particle, but since the numbers were not large enough to provide reliable statistics for separate size-groups, the all-over distribution for all particles together is shown. The general picture is very similar to the case of the quartz dust which started out charged, indicating that some kind of an equilibrium can actually be reached and is arrived at either by neutralization or by charging. It is seen that the number of singly charged particles is very little less than that of the neutral ones in agreement with the argument that $\varphi_2(1)$ can be only very little larger than $\varphi_1(0)$ for particles of this size. However, the observed curve falls off much faster than would be predicted by the simple theory indicated above. The most obvious explanation of this deviation is that ions colliding with particles do not necessarily impart their charge to the latter, especially not if the particle is already carrying an excess of the same sign as the ion. Nevertheless it is quite certain that an appreciable percentage of smoke particles in equilibrium are carrying multiple charges. The largest charge observed by us on an ammonium chloride particle of one micron diameter one day after formation was 18 electrons. It is interesting to note that this direct verification of the existence of multiply charged smoke particles in equilibrium with ionized air is a rather nice verification of the findings of G. Vassails,[7] who ascribed the appearance of groups of Langevin ions of somewhat smaller size to the multiple charges acquired from the natural ions in the air. The fact that the distributions always were practically symmetrical seems to indicate that both photoelectric effect and direct ejection of electrons by corpuscular radiation are negligible under normal conditions.

The work on dust electrification of which this is an interesting by-product was undertaken on an ONR contract at the suggestion and under the supervision of Professor L. B. Loeb. The writer desires to acknowledge his appreciation of the advice and encouragement given him in various phases of this work by Professor Loeb.

[7] G. Vassails, thesis, Masson, Paris (1948).

Double Crystal and Slit Methods in Small Angle X-Ray Scattering*

H. N. Ritland,† P. Kaesberg, and W. W. Beeman
Department of Physics, University of Wisconsin, Madison, Wisconsin
(Received February 17, 1950)

An angular resolution of about 10 seconds of arc may be obtained with a double calcite crystal spectrometer for Cu $K\alpha$-radiation. However, the shape of the rocking curve at large angles is such as to make the detection of radiation from weakly scattering materials difficult. Of the order of 10^6 counts per second are detected with a stationary, water-cooled copper target and parallel crystals. Wing intensities, even at large angles, do not fall much below 10^2 counts per second. Wing heights may be considerably reduced by the addition of a third calcite crystal. With this modification the crystal spectrometer is a useful instrument for most scattering experiments in which extremely high resolution is desired. It is inferior to a slit system and Geiger counter arrangement when poorer resolution may be tolerated but high sensitivity is required. A slit instrument with an angular resolution of 4 or 5 minutes of arc provides an effective incident flux of more than 10^6 counts per second while wing intensities may be as low as 1 count per second.

INTRODUCTION

THE study of small angle x-ray scattering involves a number of experimental difficulties in much more severe form than in conventional x-ray diffraction work. Among these are the requirements of high resolution and high sensitivity as well as good monochromatization and accurate measurement of intensity. These difficulties arise from the necessity of obtaining quantitative information about the rapidly changing scattered intensity in a narrow angular region about the primary beam.

In most small angle scattering studies higher resolution and sensitivity are obtained by making suitable refinements in photographic x-ray diffraction technique, principally involving finer slits and larger specimen-to-film spacing.[1] The diffraction chamber is usually evacuated to eliminate air scattering. Monochromatization is obtained by using radiation reflected from single crystals or by using selective filters. Accurate intensity measurements can be made only with difficulty because of the limited range of linear response of photographic film.

This paper discusses and compares the properties of two instruments which have been found useful for the study of small angle scattering, both involving Geiger counter detection. One utilizes crystals for collimation, while the other uses slits. The crystal apparatus has extremely high resolution, but has comparatively low effective sensitivity due to a background caused by diffuse scattering from the crystal surfaces. The slit apparatus has less resolution, but has high sensitivity.

The relative importance of resolution and sensitivity depends upon the scattering material being studied. A recently completed experimental study[2] of several proteins in solution illustrates the case where resolution requirements are not excessive but high sensitivity is required. Dexter and Beeman[3] describe a method for obtaining particle sizes from multiple scattering. Their method requires the resolution provided by the crystal apparatus.

TWO AND THREE CRYSTAL SPECTROMETERS

The two crystal spectrometer[4] (as well as the three crystal instrument to be described later) is shown schematically in Fig. 1. The first calcite crystal collimates the beam and is fixed in position so that line radiation from the x-ray tube strikes it at the Bragg angle. The scattering material is placed midway between the two crystals. The angular distribution of the radiation leaving the scatterer may be analyzed by rotating the second calcite crystal, since its reflectivity is high only for radiation impinging upon it at the Bragg angle. As the second crystal is rotated, that part of the scattered beam which will be reflected, does not continue to strike the same point in the crystal. As a result, the scattered radiation leaving the second crystal is focused to a point beyond it, at a distance equal to that between the second crystal and the scatterer. A thin window argon-alcohol Geiger counter is placed at this point and remains stationary as the second crystal rotates.

Slits placed at the tube window and before the sample sharply delimit the wave-length spread of the continu-

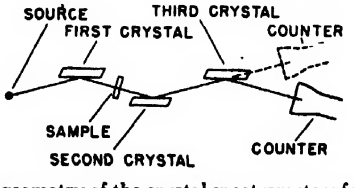


FIG. 1. The geometry of the crystal spectrometers for small angle scattering. In the double crystal arrangement only the first two crystals are used, and the counter position is shown by the dashed line drawing. The drawing has been distorted slightly in that the counter has been placed somewhat too far away from the second crystal. The full line drawing of the counter indicates its position for the three crystal arrangement.

* This work was supported in part by the joint program of the ONR and AEC and in part by the Wisconsin Alumni Research Foundation.
† Now at the University of Washington, Seattle, Washington.
[1] O. E. A. Bolduan and R. S. Bear, J. App. Phys. 20, 983 (1949).
[2] Ritland, Kaesberg, and Beeman, "An x-ray investigation of the shapes and hydrations of several protein molecules in solution" (to be published).
[3] D. L. Dexter and W. W. Beeman, Phys. Rev. 76, 1782 (1949).
[4] Kaesberg, Ritland, and Beeman, Phys. Rev. 74, 71 (1949).

ous spectrum striking the scatterer so that its effect is negligible compared with the desired line radiation. The $K\beta$-radiation may be eliminated by the proper filters. Second-order radiation is passed by the crystal and must be guarded against when strongly absorbing samples are used. However, in ordinary cases where the absorption of the line radiation is about a factor of three, second-order intensity is entirely negligible since it is only of the order of 1/1000 of the intensity of the $K\alpha$-lines at 30 kilovolts, and may be reduced further by decreasing the voltage. In addition, the Geiger counter, when filled to an argon pressure sufficient to absorb 90 percent of the first-order radiation, absorbs but 25 percent of second order.

With perfect calcite crystals and copper radiation the experimental half-width at half-maximum for a double crystal rocking curve (intensity *vs.* angle) is known to be about 5 seconds of arc. This is in good agreement with theoretical calculations.[5] Experimental rocking curves are available in the literature only for the central portion of the curve, but theory predicts that they decrease inversely as the angle squared for large angles.[6] However, it is not known to what angles this type of behavior persists in real crystals. In order to evaluate the advantage of the narrow width of the double crystal rocking curve, detailed information concerning the wing intensity at large angles must be had. For diffuse scattering the area of the central part of the rocking curve should be compared with the wing intensity, while for line scattering the significant ratio is the height of the rocking curve at the center to the wing height at the position of the scattered line.

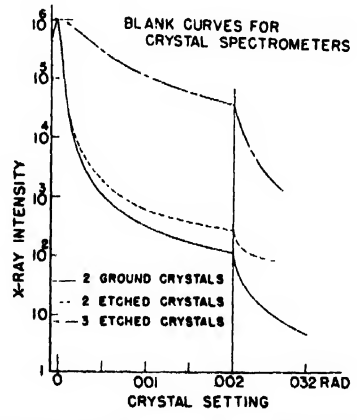


FIG. 2. The top curve is the double crystal rocking curve for a pair of calcite crystals that have been ground with carborundum to increase their reflectivity. The second curve shows the rocking curve obtained when these crystals were thoroughly etched with 0.5N hydrochloric acid. The lower curve shows the three crystal rocking curve that results when a third etched crystal is added (as described in the text). To facilitate comparison of their shapes, the three curves have been normalized to the same ordinate at zero rocking angle.

[5] A. H. Compton and S. K. Allison, *X-Rays in Theory and Experiment* (D. Van Nostrand Company, Inc., New York), second edition, pp. 709–730.
[6] F. P. Miller, Phys. Rev. **56**, 757 (1939).

TABLE I. The table shows the effect of etching treatment with 0.5 N HCl on the shape of the rocking curve for a pair of ground calcite crystals. The intensities and the areas are in arbitrary units. Copper radiation was used.

	A Etching time	*B* Center intensity	*C* Rocking curve width	*D* Rocking curve area[a]	*E* Wing height at 1°	*F* Ratio *E/B*	*G* Ratio *E/D*
1.	0 sec.	100,000	116 sec. arc	130	140	$140 \cdot 10^{-6}$	1.1
2.	5	200,000	96	220	700	350	3.2
3.	10	250,000	60	175	300	120	1.7
4.	15	240,000	30	95	90	38	1.0
5.	20	230,000	16	45	45	20	1.0
6.	25	230,000	14.5	40	20	9	0.5

[a] Within 200 seconds of arc of zero rocking angle.

We have found that curves obtained with most calcite crystals as cleaved, or ground and sufficiently etched, very nearly equal the theoretical width and indeed have wings which decrease in intensity inversely as the angle squared to about 3 or 4 minutes of arc. But the wings then go down more slowly and vary considerably with the choice of crystals and with treatment of their faces. To determine some of the effects of crystal surface conditions upon the usefulness of the instrument, a series of rocking curves was taken with a pair of ground crystals (ground with No. 600 carborundum) as a function of the time of etching in 0.5 normal hydrochloric acid. Table I summarizes the results and Fig. 2 shows rocking curves for these crystals before and after thorough etching.

It may be seen that with a light etch the central intensity is considerably increased but then remains substantially constant with further etching. The wing heights are increased at first, and then reduced as the etching proceeds, but the reduction has become appreciable only after the central area of the rocking curve is small compared with its original value. It is thus apparent that the reflected intensity may be increased by misaligning some of the surface layers, but only at the expense of a great loss in sensitivity because of the increased wing background. These results strongly suggest that best experimental conditions occur when the crystals have been etched enough to obtain as narrow a rocking curve as possible.

If a third crystal is placed so that it is always parallel to the second crystal (and the two rotated together) it should reflect almost no radiation that leaves the second crystal at other than the Bragg angle, but should reflect Bragg radiation with little diminution of intensity. Such a modification is shown schematically in Fig. 1. With this modification one no longer has a focal point at which to place the Geiger counter, so that a counter with a sufficiently large sensitive volume must be used, or else the counter must be moved as well as the second and third crystals.

The addition of a third crystal does not greatly reduce the central (i.e., all three crystals parallel) intensity of the rocking curve since the radiation has already been highly collimated by the first two crystals. We observed reductions of central intensity by a factor of between two and three upon adding a third crystal. This is a

TABLE II. A comparison of the shape of the rocking curves for two and three crystal arrangements. The curves are normalized to 1,000,000 at the center. Iron radiation was used. These curves are for well-etched, near perfect crystals.

Seconds of arc from the parallel position	Intensity 3 crystals	Intensity 2 crystals
0	1,000,000	1,000,000
25	23,000	35,000
50	5,300	8,700
100	1,000	2,300
400	100	280
1,000	33	150
4,000	8	110
8,000	3	100

slightly smaller percent reflection than theory would predict for perfect calcite crystals.

However one expects the wings to be reduced about twice as much as the central intensity with the addition of a third crystal parallel to the second. This results from the fact that when the first two crystals are not parallel the radiation leaving the face of the second can be divided into two equal parts one of which is strongly and the other very weakly reflected by the third crystal. The two parts are a component reflected at the center of the single crystal diffraction pattern by the first crystal but at the wing by second crystal and a component reflected on the wing by the first crystal but the center by the second. Only the latter component is appreciably reflected by the third crystal. It is diminished by the third reflection by about the same factor as is the central intensity when all three crystals are parallel.

The three crystal rocking curve also should have wings diminishing as the reciprocal of the square of the angle from parallelism.

A glance at Table II shows that the three crystal wings go down more slowly than the predicted $1/\varphi^2$ at large angles but that the improvement on the two crystal arrangement may still be much greater than a factor of two. At the largest angles the three crystal wings are less than the two crystal by more than an order of magnitude. This implies that no appreciable fraction of the wing intensity above the theoretical $1/\varphi^2$ contribution leaves the second crystal at the Bragg angle.

The cause of the excess wing intensity is not clear. One of the components leaving the second crystal is an intense beam from the first crystal which is presumably only weakly reflected by the second. This beam may be the source of temperature diffuse scattering at the second crystal or of a broad scattering peak from very small crystalline domains.

Although the third crystal has a pronounced effect in reducing wing heights, they are still about 10 times room background for large wing angles (with a stationary copper target tube). Thus the sensitivity of the three crystal spectrometer, while higher than the two crystal instrument, is still inadequate for diffuse scattering studies with materials of low scattering power. Since the scattered intensity depends upon the square of the difference in electron density between the particles and their surrounding medium, this lack of sensitivity will exclude the important class of experiments in which the scattering specimens are biological materials in aqueous solution.

SLIT COLLIMATION APPARATUS

If slits, rather than crystals, are used for collimation, resolution, as well as intensity, is governed by the slit widths and separations. As a result, for a given distance between two collimating slits, the highest attainable resolution is determined by the lowest tolerable intensity. Thus, the Geiger counter, while not an essential feature of a slit collimating system (as it is in the crystal spectrometers), may be used to advantage because it has high efficiency for detecting radiation.

The instrument described in the succeeding paragraphs was designed to study the angular distribution of scattered x-rays from particles in the size range 20A to 200A in diameter. The slit geometry was such that a convenient level of scattered intensity could be maintained with weakly scattering protein solutions as scattering specimens and with a water-cooled, stationary copper or iron target, x-ray tube as the radiation source.

The geometry of the slit system is indicated schematically in Fig. 3. A pair of stationary tantalum slits collimates the beam incident upon the scatterer. A second pair of slits analyzes the angular distribution of the radiation leaving the scatterer. The second pair of

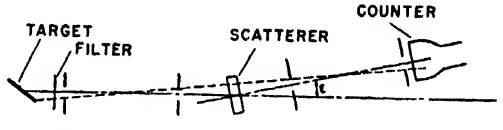


FIG. 3. A schematic drawing of the slit system described in the text. The scale has been greatly compressed in the horizontal direction. The angle ϵ is the smallest rocking angle at which the counter no longer sees the target.

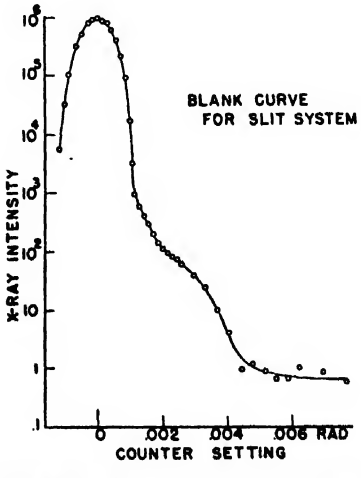


FIG. 4. The blank rocking curve for the slit system described in the text. The curve shown is for slit widths of 0.02 cm. The $4d/a$ angle is then 0.0027 radian.

slits, and also the specimen holder and the Geiger counter, are fixed to an arm which may be rotated, by means of a calibrated screw, about an axis through the center of the specimen. The distance between successive slits is 30 cm and the scatterer is halfway between the second and third slits. The slits are 1 cm high, but can be cut down by suitable apertures. The slit widths are adjustable, usually being between 0.1 and 0.4 mm wide. In order to reduce the loss by air absorption, the slits are mounted in two evacuated tubes with windows of 0.001-in. Nylon film at each end. Monochromatization is achieved by means of the Ross filter method. The differential filters are mounted on a slide in front of the x-ray tube exit window. The scatterer, Geiger counter, and filters are not in the evacuated region.

If intensity *vs.* angle is plotted for the slit apparatus with no scattering sample, a curve is obtained that is analogous to the rocking curve for two and three crystal spectrometers. This curve is shown in Fig. 4. The central part of the curve consists mainly of radiation that goes straight through all of the slits. Wing radiation arises from slit edge scattering, and from air scattering in the non-evacuated regions. If the slit width is d and the separation is a, the counter no longer sees the x-ray tube target at an angle of rotation of $2d/a$ from the straight through position (as shown in Fig. 3). Scattering from the edges of the second and third slits is eliminated at an angle of $4d/a$. Air scattering from the non-evacuated region around the sample falls off rapidly with angle, because the region where the two telescopes overlap decreases rapidly.

The most striking feature of the slit rocking curve (Fig. 4) when compared with the crystal rocking curves (Fig. 2) is its exceedingly low wing intensity. At angles only slightly greater than $4d/a$, the ratio of center to wing intensity is greater than 10^6. The three crystal rocking curve approaches this ratio only at comparatively large angles (see Fig. 2) while the two crystal curves are two or three orders of magnitude worse.

Under normal operating conditions a slit instrument with an angular resolution of 4 or 5 minutes of arc provides an effective incident flux of about 10^6 counts per second. This is approximately the counting rate for the three crystal spectrometer at zero rocking angle. However, a more significant comparison for diffuse scattering is the area under the central portion of the rocking curve for each instrument. The slit system curve area exceeds that of the three crystal instrument by a factor of 25.

We wish to thank the Kimberly-Clark Corporation for their support of the Kimberly-Clark Fellowship held by one of us (H.N.R.) during the year 1948–49.

Letters to the Editor

Vacuum Tubes with Mutually Bombarding Oxide Cathodes

E. G. Hopkins

Electrical Engineering Department, The New South Wales University of Technology, Sydney, New South Wales, Australia

March 13, 1950

EXPERIMENTS in which two flat oxide cathode surfaces 3.5 cm long and 1.0 cm wide were mounted face to face to form a high vacuum diode and were caused to act alternately as anode and cathode in passing a current of 50 cycles per second between them have shown that such an arrangement can be made to operate stably after the removal of the initial heating source. If a 240-volt a.c. supply is connected to the cathodes through a series resistance at an instant when the temperature of their hottest parts is above a critical minimum (below the visible range), the current builds up to a value determined by the space-charge limited characteristics of the diode. One diode with a spacing of 3 mm was found to obey closely the law, r.m.s. amp. $=1.9$ (r.m.s. volts)$^{1.42}\times10^{-4}$ between 120 and 220 volts, the corresponding range of maximum cathode brightness temperatures measured on an uncoated nickel spot being 880 to 1260°K. The current wave form was quite symmetrical about the zero axis and showed no signs of saturation. Portions of the cathode below a visible temperature glowed with a pale blue fluorescence under bombardment. One sealed-off diode was allowed to run for seven hundred hours.

A triode was constructed consisting of two cathodes as described having a total separation of approximately 2 mm and a grid of molybdenum wires spaced equally between them. Tests were made with a grid voltage applied through a synchronous switch in such a way that for any half-cycle it consisted of a half-sine wave applied negatively relative to the particular electrode then acting as a cathode. By assuming the Langmuir equation for triode current to hold for all instantaneous voltages, the equation I_p (r.m.s.) $= K$ (E_b r.m.s. $+ \mu E_g$ r.m.s.)^n was used to estimate μ, the amplification factor from values of voltage and current read on r.m.s. reading instruments. This equation is useful when examining the possible use of such a triode to control a.c. power circuits.

The triode was made to oscillate at a frequency of 0.5 Mc/sec. in an oscillator circuit which included a synchronous switch to reverse the phase of the grid drive. When operating stably with a cathode brightness temperature of about 950°K, the oscillator drew 0.12 amp. at 220 volts and delivered approximately 5 watts to the load lamp.

These experiments demonstrate the possibility of utilizing the power usually wasted in anode dissipation to heat a cathode surface much larger than can be provided economically in conventional power tubes. It appears feasible from the data obtained from this work and from experience with conventional tubes that if the spacings of a practical tube based on the above principle were made as small as possible without causing excessive grid emission, hot spot formation, or mechanical troubles, then such a tube could be made to deliver useful amounts of high frequency power at acceptable efficiency without the use of transformers or rectifiers when operated directly from an a.c. supply of 240 volts or over.

These experiments, details of which will be published when the present work has reached a more advanced stage, were carried out by the writer at the Radiophysics Vacuum Laboratory in the Electrical Engineering School of the University of Sydney.

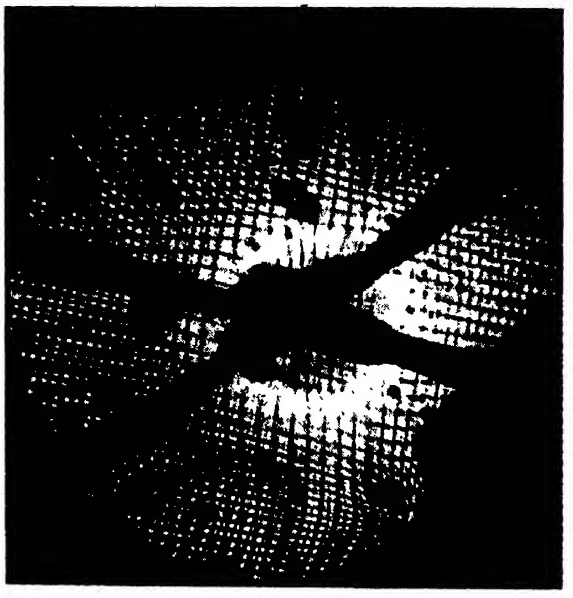


FIG. 1.

Electron Optical Properties of Space-Charge Clouds

L. MARTON AND D. L. REVERDIN
National Bureau of Standards, Washington, D. C.
April 13, 1950

THE space-charge lens has been repeatedly considered as a corrective element for the reduction of the aperture defect of electron lenses.[1] This is due to the recognition by Scherzer[2] that no axially symmetrical, time invariant, space-charge free field is free from the aperture defect, and that dropping of any of these conditions may improve the behavior of the lens. Gabor,[3] who was one of the first to propose the use of space-charge clouds as corrective elements, raised, however, a question and suggested an experiment: "Will the 'space-charge' cloud not act on the beam like frosted glass?" He proposed "to settle this point by the simple experiment of shooting a fast electron beam parallel to the axis through a static magnetron. It may be mentioned that such an experiment would also at once settle the question how nearly Hull's simple theory agrees with actual conditions in a magnetron as, on Hull's theory, the beam would suffer no radial deflection."

Gabor's proposed experiment forms the subject of a dissertation for a Ph.D. degree by one of us[4] and will be published separately. The present brief communication deals with a partial answer to Gabor's question about the suitability of the space-charge cloud as a lens-correcting element.

In the present experiment, a wide electron beam was sent axially through a static magnetron. Before reaching the magnetron, the beam passed one wire mesh, and a second wire mesh was placed in the beam's passage after the magnetron. The details of the experimental conditions are given in reference 4. A typical observation is shown in Fig. 1. This figure represents the photographed shadow patterns cast on a fluorescent screen. The shadows of the two wire meshes are plainly visible and in addition to them, there are two heavier shadows shown at approximately a 60° angle. These latter are the shadows of the filament supporting bars. Very marked in Fig. 1, also, is the bright ring located between the cathode (in the center of the picture) and the anode (dark circular shadow at the circumference of the picture). The outer edge of this bright ring lies approximately where the theoretical electron cloud radius would be according to the classical magnetron theory.

Approximate computations carried out on this pattern (Fig. 1), show that the bright ring is produced by a high intensity space-charge cloud acting as a divergent lens. Although experimental conditions prevented us from actually measuring the focal distance of this divergent lens, this latter could be estimated to be rather large. A comparison of the pattern shown in Fig. 1, with patterns taken in absence of the space charge, shows that a partial correction of the aperture defect is probably present, although we do not have quite definite indications due to the preliminary nature of this investigation.

The wire shadows in the bright region of the picture appear much less sharp than toward the periphery. When placing an oscilloscope in the plate circuit of the magnetron, we could observe that whenever the electron optical observation showed the presence of space charge, considerable random noise could be observed in the plate circuit. Although our experimental set-up does not exclude entirely the possibility of very high frequency oscillations, the probability of such oscillations being present is extremely reduced. Thus the lack of definition in the bright region can be fairly well correlated with the noise and the conclusion drawn that the image definition is reduced by random noise in the space charge.

The present experiments thus seem to indicate that, whereas a correction of the aperture defect may be possible by means of space charge, Gabor's suspicion of its acting as a "frosted glass" seems to be justified. The present observations, being mostly an outgrowth of the magnetron investigation, are not presented as arguments for abandoning all attempts toward correction of aperture defect by means of a space-charge cloud. They seem to show, however, that the reduction of the resolution in the image may outweigh considerably whatever advantage a space charge could offer as a lens-correcting element.

[1] V. E. Cosslett, *Electron Optics* (Clarendon Press, Oxford, 1946), pp. 142–147.
[2] O. Scherzer, Optik 2, 114 (1947).
[3] D. Gabor, *The Electron Microscope* (Chemical Publishing Company, Inc., New York, 1948), p. 124.
[4] Doctoral dissertation submitted by D. L. Reverdin in January, 1950, to George Washington University, Washington, D. C. (to be submitted in abbreviated version to J. App. Phys.).

Journal
of
Applied Physics

Volume 21, Number 9 September, 1950

The Thickness Measurement of Thin Films by Multiple Beam Interferometry*

G. D. Scott, T. A. McLauchlan, and R. S. Sennett
Department of Physics, University of Toronto, Toronto, Canada

(Received February 2, 1950)

A method is described for the measurement of the thickness of thin films using the multiple beam interference techniques. Only simple equipment is required, but reduction of the order of interference to as low as *one* makes possible the measurement of average thicknesses down to 10A. Applications have been made to evaporated films of six different metals and the method appears to be useful for all stable thin films in thicknesses up to several microns.

Possible errors in the method and the precautions which may be taken to avoid them are discussed.

INTRODUCTION

VARIOUS applications of multiple beam interferometry to the study of surface topography have been discussed by Tolansky.[1] One of the important applications is to the measurement of the thickness of thin films by determination of the height of the step made by the edge of the film when formed on an optically smooth substrate. The method involves covering the film and adjacent substrate with a highly reflecting coating and then measuring the resulting step in the coating. Donaldson and Khamsavi[2] have shown that for silver films with a silver coating this procedure gives the true thickness of the film. However, Avery[3] has noted that errors may be introduced in some cases, for example when chromium is used as the reflection coating for measuring a silver film.

Reflection fringes with white light as viewed in a spectroscope have been found to be the most satisfactory type of fringes for such measurements. Reflection rather than transmission fringes are used since then the reflection coating can be made sufficiently thick to eliminate a possible difference in the phase change upon reflection at the coating which is on the film compared to that which is on the substrate alone. Silver is the best metal for the coatings since it has the highest reflection coefficient, at least in the most convenient region of the spectrum. White light and a spectroscope are employed rather than monochromatic light since then optical flats are not required for the substrates and all measurements can be made simply from the wave-length drum of the spectroscope.

EXPERIMENTAL

The general arrangement of the equipment is shown in Fig. 1. The interference plates are mounted on the table of a small goniometer, the arms of which carry the source and the projector lens, L_2. Light from the source, which was a 6-volt, 32-cp incandescent lamp, after passing through the condensing lens L_1 is reflected normally to the interference plates by a half-silvered mirror. An image of the plates is formed on the slit of the spectroscope by the projector lens. The focusing of the condensing lens is not critical, but a convenient adjustment is to have an image of the source formed at the aperture of lens L_2. To ensure sharp fringes the effective angular aperture of the projector lens should be less than 1/15. The spectroscope was a Hilger constant deviation instrument with calibrated wave-length drum. A 12× eyepiece with single vertical cross hair was employed. The projector lens was a 32-mm microscope objective which produced an image of the interference plates enlarged 3× at the slit of the spectroscope. Thus the over-all lateral magnification was 36×.

The details of the interference plates are shown in Fig. 2. Standard 1 in.×3 in. microscope slides were used for these plates. The front slide is coated with silver to give a 1 percent to 2 percent transmission; the back slide is the substrate for the thin film to be measured. A step or channel is made in the film along the length of the

* Presented in part at the annual meeting of the E.M.S.A., at Washington (October, 1949).

[1] S. Tolansky, *Multiple Beam Interferometry* (Clarendon Press, Oxford, 1948).

[2] W. K. Donaldson and A. Khamsavi, Nature 159, 228 (1947).

[3] D. G. Avery, Nature 163, 916 (1949).

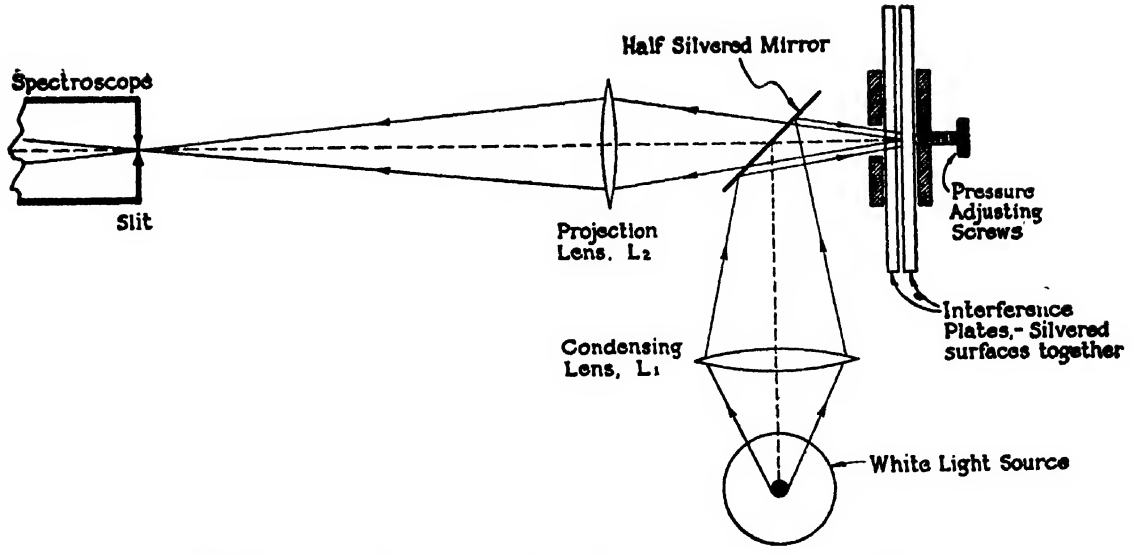


FIG. 1. Experimental arrangement for producing multiple beam interference fringes.

slide either during or after formation. The slide is then heavily coated with silver by thermal evaporation in vacuum. About 1000A is sufficient to eliminate the possible phase differences mentioned above. This silver surface thus has a step or channel, the depth of which corresponds to the thickness of the original film. It is important that the silver coatings have low optical absorption. In addition to the usual conditions for obtaining such coatings, it has been found that a rapid rate of deposition of the silver is also necessary.[4] The mounting which holds the plates on the goniometer table has a 3-mm aperture in the front and two small pressure-adjusting screws at the back. These screws control the separation of the plates.

The step or channel can be made in any convenient manner. It is desirable of course to have as sharp edges as possible. For metallic films, one method which was found quite satisfactory was to form a channel by gently drawing a not-too-sharp needle across the film. Some experience is required in order to remove all the metal yet not to scratch the glass substrate. The edges of the channel should be sharply defined and the bottom quite smooth. The slightest marring of the glass is readily detectable from the shape of the fringes. Another method is to form the film with the step or channel by using a suitable shield such as a narrow strip of metal foil. Both methods were found to give the same thickness measurements within the limits of error.

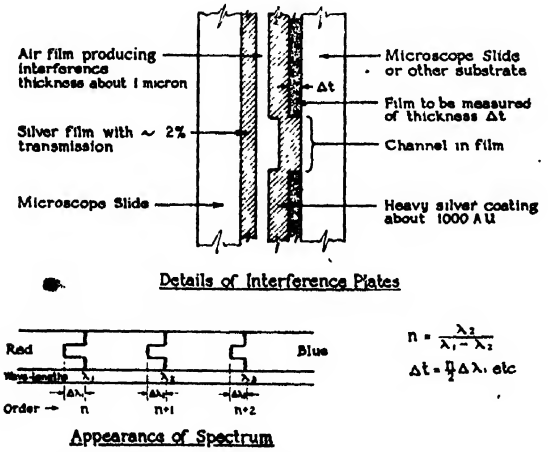


FIG. 2. Details of the coatings on the interference plates and a schematic diagram of the fringes from the plates.

A number of adjustments are required to obtain satisfactory fringes. The spectroscope is focused for monochromatic light and the slit width set for maximum resolution. The components as shown in Fig. 1 are aligned geometrically so that when the plates are in position a white light spectrum with interference fringes is observed in the spectroscope. Adjustments are made as follows.

1. Lens L_2 is focused to give sharp fringes.
2. The pressure screws are adjusted to make the fringes nearly vertical, at least in the region of the step.
3. Lens L_2 is moved laterally to locate the position where the fringes remain stationary. This corresponds to the area where the plates are exactly parallel and hence where the fringes have optimum sharpness and contrast.
4. The pressure screws are then set for the desired order and at the same time adjustment (2) is maintained, and (3) is checked.

While moving lens L_2 as in (3), fringes are viewed corresponding to various positions along the step. In this way local irregularities can be noted and avoided so that a final representative position is selected.

A schematic diagram of the spectrum observed is given in Fig. 2. The wave-length interval $\Delta\lambda$ corresponding to the height of the step Δt is measured directly, using the wave-length drum of the spectro-

[4] R. S. Sennett and G. D. Scott, J. Opt. Soc. Am. 40, 203 (1950).

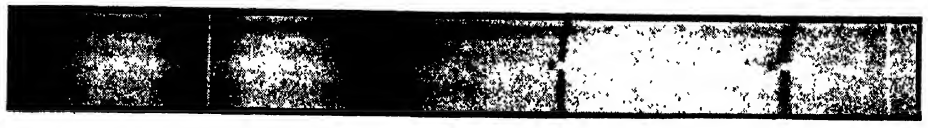


FIG. 3. Spectrogram of multiple beam interference fringes of orders 12 to 16. The mercury lines superimposed as wave-length references are the yellow doublet at 5770A and 5791A, the green line at 5461A, and the blue line at 4358A. The discontinuity in the fringe is caused by a channel in a silver film 115A thick.

scope. The order n for the fringe is determined as indicated by the formula in Fig. 2 and then Δt may be calculated: $\Delta t = (n/2)\Delta\lambda$. The spectrum of a typical fringe system at relatively high orders, *viz.*: 12 to 16, is shown in Fig. 3. Measurements were usually made at orders from 1 to 6. At order *one* for example, the "discontinuity" in the fringes of Fig. 3 appears as shown in Fig. 4. The mercury spectrum has been superimposed on these spectrograms merely to indicate the wave-length scale.

Since microscope slides are used in place of optical flats, the fringes are not in general straight; however, the use of the slides, which are relatively thin, makes it possible to obtain interference orders as low as *one* or *two* quite readily. Such low orders add greatly to the convenience with which readings can be made. It may be noted, for example, that for order *two* the wave-length interval at the step is just equal to the thickness of the film. If the fringe for order *two* is at the red end of the visible spectrum, the fringe of order *three* is in the blue, so that the order number can be immediately recognized. Furthermore it was observed that with microscope slides the fringes were about one-quarter the width of those obtained with ordinary optical flats. This presumably indicates that over distances of say 10 microns the microscope slides which are rolled or fire polished are smoother than the flats which are polished by an abrasive. Irregularities of 10 to 20A in the flats would account for the observed effect and it seems possible that these could be present in standard quality flats.

The fringes are produced by a very narrow area of the interference plates—the area corresponding to the image of the slit at the plates. With the arrangement described here, an individual setting on a fringe represents a position which averages heights over an area of about 5×50 microns on the plates. However, because of the necessary adjustments and the desirability of avoiding local irregularities, it is preferable to have available a film area of at least the order of 1 sq. mm.

It was found that within experimental error measurements of the depth of a step or channel were the same at all useful orders, i.e., from *one* to about *thirty*. For interference order *two* the wave-length settings on a fringe could be made in general to within 2A. Hence for a film 100A thick the measurement should be correct certainly within 5 percent.

A possible source of error in these measurements is the variation with wave-length in the phase change for reflection at a thick silver film. From a plot of $n\lambda$ against λ, it was found that this variation is quite small. The approximate correction for this effect in the measurement of the thickness Δt is given by: $\Delta t = (n/2)\Delta\lambda + [0.01(3)/2]\Delta\lambda$. It is evident that even at order *one* the correction is only ½ percent and hence for the sharpness of fringes at present obtainable this error can be neglected.

FIG. 4. The same channel as in Fig. 3, but at order *one*. The mercury yellow doublet and green line are superimposed on the spectrum.

As Tolansky[1] has emphasized, the method is capable of high resolution in the measurement of average heights. This fact is illustrated in Fig. 5 which shows a fringe at order *one* used to measure a silver film 15A in thickness.

DISCUSSION

The displacement of the discontinuity in the fringe certainly gives the depth of the channel or step in the silver coating, but whether this is precisely the same as the thickness of the original film must be considered in detail. It seems reasonable to assume that it is in the case of films which are continuous and of much the same nature as the glass substrate, such as Formvar films. However for films which are either aggregated or quite different in nature from the glass substrate, the assumption may not hold. Avery[3] has shown that when a chromium coating is used to measure silver films, the apparent thickness is about 30 percent greater than when silver is used as the coating. This may be explained by assuming that chromium does not fill the interstices of a silver film as completely as does silver itself. Alternately, it could also be assumed that the metal just adjacent to the glass is less aggregated for the chromium than for the silver. However, Donaldson and Khamsavi[2]

FIG. 5. The discontinuity in a fringe due to a silver film of average thickness 15A.

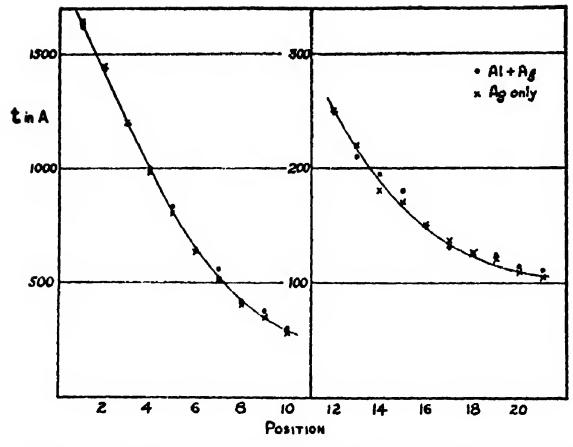


FIG. 6. Effect of an intermediate layer of aluminum on aluminum film thickness measurements using silver for the high reflection mirror.

have shown that when a silver coating is used for a silver film the measured thickness gives a density for the metal in the film equal to that of the bulk metal. This result is applied to films as thin as about 100A, which are known to be aggregated. Since electron diffraction shows that lattice spacings in thin films are essentially the same as in the corresponding bulk metals, it can be concluded that at least for the case of silver the multiple interference method yields the true average thickness for aggregated as well as continuous films.

For films other than silver the following method can be used in order to avoid errors due to the effects mentioned above. The film to be measured, A, is first covered with an intermediate layer, B, of the same material and deposited under the same conditions as A. The coating B is made sufficiently thick so that its top surface has the same structure over the substrate as over the film A. Finally the high reflection silver coating is added. It is evident that this procedure compensates for peculiarities (1) in the form of the material of the film in contact with the substrate and (2) the way in which the first layer of the silver coating may be deposited on the film material.

Thin films of the following metals, in addition to silver, have been measured in the range from 40A to 1500A: gold, copper, aluminium, antimony, and chromium. For each case at least two channels were made side by side through the metal films which varied in thickness along the length of the microscope slides. One channel was first coated with a thick film of the same metal, and then both were coated with a heavy silver film. For all the metals studied the thickness measurements were the same within the limits of error, either with or without the intermediate layer. As an example the results for aluminium are shown in Fig. 6. The measurements for the case of silver on chromium showed no discrepancy though Avery found that errors do occur when chromium is used as the coating for silver. It can be concluded that when silver is used as the reflection coating, the use of the intermediate layer is unnecessary for films of gold, copper, aluminium, antimony, and chromium.

It is essential to the method that the intermediate coating, or in the simpler case the silver coating, be of precisely the same thickness on the bare substrate as on the film. This condition is readily obtained when the films are deposited by thermal evaporation in vacuum, provided none of the incident atoms are reflected or scattered away from the surface which they strike. It is known that the distance of atomic migration over the substrate is much too small to be of importance in this connection. However, during the initial stages of the formation of these coatings, differences in the scattering of atoms from the films compared to the scattering from the substrates could occur. It is highly probable that this effect would only be important for films less than 50A in thickness. Because of the difficulties of making independent and accurate measurement of films of such a thickness, no experimental information on the importance of the effect is available. The possibility of such an error could be eliminated by forming the film on a substrate of the same material as the film itself.

CONCLUDING REMARKS

The method described here for the measurement of the thickness of thin films appears applicable to any stable thin film which can be formed on a smooth substrate and with a sharp edge. It is particularly suitable to film thicknesses in the range from 10A to 1000A, but can be used for films up to several microns in thickness. The precision at 100A is better than 5 percent.

The equipment required is simple and the results can be obtained quickly. After the film has been properly coated, about two minutes is usually sufficient time to make the adjustments and obtain a thickness measurement.

There are certain possible errors in the method, but in many cases these are not important. However each particular type of film and substrate should be considered independently to determine what precautions may be necessary.

Distribution in Energy of Johnson Noise Pulses

BEN R. GOSSICK*
NEPA Division, Fairchild Engine and Airplane Corporation, Oak Ridge, Tennessee
(Received August 22, 1949)

The distribution in energy of Johnson noise pulses (random thermal voltages in a resistor) is derived and checked against pulse-height measurements made with a linear amplifier and electronic counter. The following parameters are determined from the distribution function: (1) time of flight associated with a Lorentz mean free path, (2) the number of electrons which produce a pulse of a given height, and (3) the potential an electron falls through in a Lorentz mean free path.

DISCUSSION

NYQUIST'S equation[1]

$$d\langle V^2\rangle = 4kTRdf \tag{1}$$

applies to random thermal voltages across either an ohmic resistor or the radiation resistance of a Hertzian dipole. Bernamont,[2] and later Bakker and Heller,[3] using Lorentz electron theory and Fermi-Dirac statistics, obtained a proof of Eq. (1) in which the ohmic resistance appears explicitly. The distribution of random noise currents has been treated by Landon,[4,5] Norton,[5] Kac,[6] and Rice.[7] Norton draws upon Rayleigh's[8] solution for the amplitude distribution of a large number of unit vibrations with random phase.

The purpose of this paper is twofold: one, to present a solution for the Johnson[9] noise pulses by a method which is thought to be new, and two, to supply measurements of Johnson noise pulses made with a linear pulse amplifier and electronic counter. It is believed that such measurements have not been previously reported.

The distribution of noise pulses is analyzed here by Maxwell-Boltzmann statistics using a time-energy phase space. It is assumed that the resistor is in thermal equilibrium, and that the Johnson noise is a stationary random process. In considering the distribution of pulses by Maxwell-Boltzmann statistics, it is convenient to employ the familiar treatment of molecules as an analog. This analogy is given in Table I.

The probability of a distribution of pulses $p_1, p_2 \cdots$ becomes

$$W = p! \prod_i (g_i)^{p_i}/p_i!. \tag{2}$$

As the number of pulses is large, Stirling's approximation is applicable, so that (2) can be rewritten as

$$W = (p)^p \prod_i (g_i/p_i)^{p_i}. \tag{3}$$

In order to obtain the most probable complexion, $\ln W$ is maximized by the method of Lagrangian multipliers. The following equations of condition are adopted:

$$\sum_i p_i - (t'/\tau') = 0, \quad \tau' \geqslant \tau, \tag{4}$$

$$\sum_i p_i - (t'/\tau) = 0, \quad \tau' < \tau. \tag{5}$$

$$\sum_i p_i E_i - E = 0, \tag{6}$$

where τ is the time of flight associated with the Lorentz mean free path, τ' is the response time of the electronic system, E_i is the energy of a pulse in the ith cell of phase space, and E is the total energy of all pulses occurring in the interval $0<t<t'$. Equations (4) and (5) state that the total number of pulses in the counting interval $0<t<t'$ is invariant, and this is an approximation requiring that $t' \gg \tau$, and $t' \gg \tau'$. Because of the limitations of electronics, Eq. (4) is normally applicable. Equation (5) is included for completeness to the theory, and because it seems possible that it could be used in special cases, e.g., for τ across a biased barrier of a semiconductor-metal rectifier. Equation (6) follows from the original assumption that the resistor is in thermal equilibrium. Solving for the number of pulses p_i in the ith cell of phase space yields:

$$p_i = g_i e^{-1-\alpha-\beta E_i}, \tag{7}$$

where β and α are the Lagrangian multipliers associated with (6) and (4), respectively. Following the molecular analog, (7) can be expressed as the differential:

$$dp^2 = Ae^{-\beta E} dE dt, \tag{8}$$

which gives the number of pulses with energy between E and $E+dE$ arriving between t and $t+dt$, where A is a constant to be determined directly.

TABLE I.

Molecules	Pulses
1. Container of volume V enclosing M molecules	1. Interval of duration t' enclosing p pulses
2. Dimensions of box are large compared to dimensions of molecules	2. Interval is large compared to pulse width
3. Statistical weight is proportional to $\Pi_i dx_i dp_i$	3. Statistical weight is proportional to $dEdt$

* Present address: Physics Department, Purdue University, Lafayette, Indiana.

[1] H. Nyquist, Phys. Rev. 32, 110 (1928).
[2] H. J. Bernamont, Ann. de physique 71 (1937).
[3] C. J. Bakker and G. Heller, Physica VI, 262 (1939).
[4] V. D. Landon, Proc. I.R.E. 29, 50 (1941).
[5] V. D. Landon and K. A. Norton, Proc. I.R.E. 30, 425 (1942).
[6] M. Kac, Am. J. Math. 65, 609 (1943).
[7] S. O. Rice, Bell Sys. Tech. J. 23, 282 (1944); 24, 46 (1945).
[8] Lord Rayleigh, *Theory of Sound* (Dover Publications, New York, 1945), Chapter II, 42a.
[9] J. B. Johnson, Phys. Rev. 32, 97 (1928).

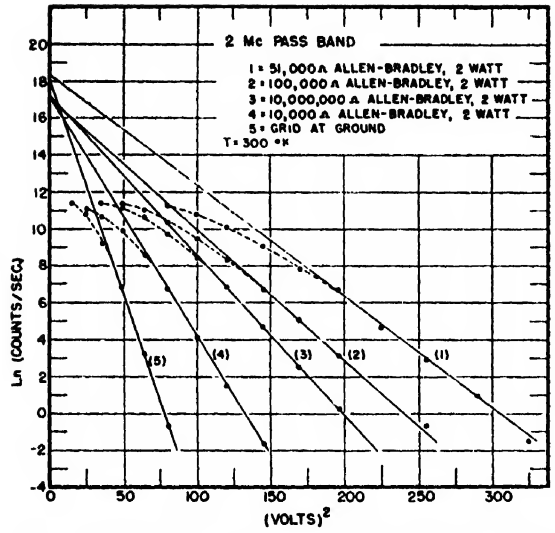


FIG. 1. (Volts)² refers to the square of the pulse-height selector setting at the A-1 amplifier. Hence, to relate these curves to (23) it is necessary to introduce the over-all gain G into (23) as follows:

$$p_{V_0} = (1/\tau') \exp(-V_0^2G^2/2\langle V^2\rangle G^2)$$
$$= (1/\tau') \exp[-(\text{volts})^2/2 \text{ (mean square volts)}]$$
$$\ln p_{V_0} = \ln(\text{count/sec.})$$
$$= -(\text{volts})^2/[2 \text{ (mean square volts)}] + \ln 1/\tau'.$$

The observable energy per pulse equals

$$E = V^2\tau'/2R, \tag{9}$$

where V is the amplitude of the voltage pulse. By means of (9) Eq. (8) can now be rewritten in terms of voltage as

$$dp^2 = (A\tau'V/R)dVdt \exp(-\beta\tau'V^2/2R). \tag{10}$$

As the total number of pulses which arrive in the interval t' is the quotient t'/τ', it follows that

$$t'/\tau' = \frac{A\tau'}{R}\int_0^{t'}\int_0^{\infty} VdVdt \exp\left(\frac{-\beta\tau'V^2}{2R}\right). \tag{11}$$

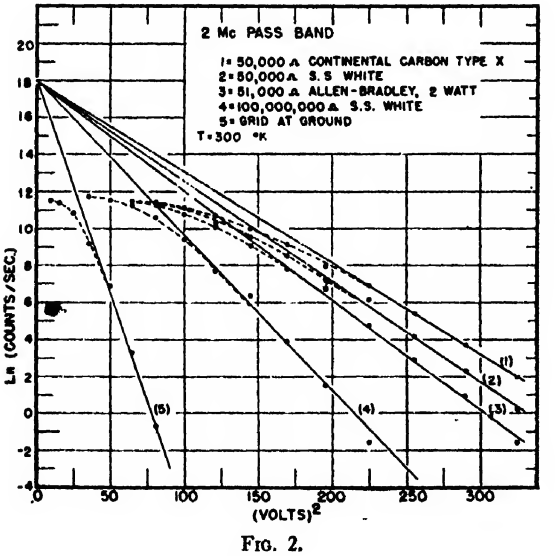


FIG. 2.

Solving (11) for A yields

$$A = \beta/\tau', \tag{12}$$

which can be used to rewrite (10) as

$$dp^2 = (\beta V/R)dVdt \exp(-\beta\tau'V^2/2R). \tag{13}$$

As it was initially assumed that the Johnson noise is a stationary random process, the function of interest is the mean pulse rate, hereafter denoted by $\dot{p}$. Thus

$$d\dot{p} = (\beta/R)VdV \exp(-\beta\tau'V^2/2R) \tag{14}$$

gives the average rate of pulses with amplitude between V and $V+dV$.

The constant β is evaluated immediately in terms of the mean power $\bar{P}$.

$$\bar{P} = \frac{\beta}{R}\int_0^{\infty} EVdV \exp\left(\frac{-\beta\tau'V^2}{2R}\right), \tag{15}$$

$$= \frac{\beta}{R}\int_0^{\infty} \left(\frac{V^2\tau'}{2R}\right)VdV \exp\left(\frac{-\beta\tau'V^2}{2R}\right), \tag{16}$$

$$= \frac{\beta\tau'}{2R^2}\int_0^{\infty} V^3dV \exp\left(\frac{-\beta\tau'V^2}{2R}\right), \tag{17}$$

$$\bar{P} = 1/\beta\tau'. \tag{18}$$

$$\beta = 1/\bar{P}\tau', \tag{19}$$

$$= R/\langle V^2\rangle\tau'. \tag{20}$$

Equation (13) is now rewritten, substituting for β the value just obtained in (20),

$$d\dot{p} = (VdV/\tau'\langle V^2\rangle) \exp(-V^2/2\langle V^2\rangle). \tag{21}$$

A convenient function to obtain experimentally is the average counting rate of the pulses which have a height exceeding the bias setting of a discriminator circuit. If the bias setting is V_0, it follows that the rate $\dot{p}$ should equal

$$\dot{p}_{V_0} = \frac{1}{\tau'\langle V^2\rangle}\int_{V_0}^{\infty} VdV \exp\left(\frac{-V^2}{2\langle V^2\rangle}\right), \tag{22}$$

$$= (1/\tau') \exp(-V_0^2/2\langle V^2\rangle). \tag{23}$$

Results which are essentially the same as (23) can be found in the excellent mathematical analyses of the noise current envelope by Kac[6] and Rice.[7] In order to demonstrate this agreement, it is necessary to apply their results to the number of times per second the noise voltage is expected to pass through V_0 with a positive slope, as seen on the output of a broad band linear amplifier.

Although the Lorentz mean collision time is not directly observable when $\tau' \gtrsim \tau$, the value of τ can be calculated by using (21) together with a random walk solution after Wang and Uhlenbeck.[10] These authors

[10] Ming Chen Wang and G. E. Uhlenbeck, Rev. Mod. Phys. 17, 323 (1945).

have shown that for a one-dimensional model, the probability of a resultant displacement $\nu = n - m$ out of a total of S unit random steps equals

$$P(n|m,S=)\frac{S!}{\left(\frac{\nu+S}{2}\right)!\left(\frac{\nu-S}{2}\right)!}(\tfrac{1}{2})^S. \quad (24)$$

If an assumption is made that $S \gg \nu$, and Sterling's approximation is applied, the following relation is obtained from (24),

$$P_{\nu,S} \propto \exp(-\nu^2/2S). \quad (25)$$

The number of random electron displacements S during an interval of observation τ' is given by

$$S = N\tau'/\tau, \quad (26)$$

where N is the total number of conduction electrons in the resistor. By combining (25) and (26) one obtains the following relation for the probability that ν-electrons will contribute to a pulse during the interval of observation τ'.

$$P_{\nu,\tau'} \propto \exp(-\nu^2\tau/2N\tau'). \quad (27)$$

As (21) and (27) are associated with the same event, it is permissable to equate the exponents of the two expressions. Thus

$$\nu^2\tau/2N\tau' = V^2/2\langle V^2\rangle. \quad (28)$$

It is necessary now to obtain the relationship between the pulse amplitude V and the number of electrons ν which contribute to the pulse in order to evaluate τ in (28).

The pulse amplitude V can be expressed by Ohm's law as the product of the peak noise current I (based on ν electrons) and the constant resistance R (based on the total number of conduction electrons N). Thus the current I and resistance R can be generally expressed as

$$I = \nu\epsilon\lambda/L\tau'\sqrt{3}, \quad (29)$$

and

$$R = L^2/N\epsilon\mu, \quad (30)$$

where ϵ is the electronic charge, λ is the Lorentz mean free path, L is the linear length of the resistor, and μ is the electronic mobility. A few simple steps of algebra applied to (28)–(30) yields

$$\tau = \epsilon\lambda^2 R/(3\langle V^2\rangle\tau'\mu). \quad (31)$$

Analyses of the mean square noise voltage $\langle V^2\rangle$ by Bernamont[2] and Moulin[11] show that a resistor and its distributed capacitance form a system in which the equipartition principle applies. Thus, $\langle V^2\rangle$ can be expressed as

$$\langle V^2\rangle\tau'/2R = C\langle V^2\rangle/2 = kT/2, \quad (32)$$

where C is the distributed capacitance.

[11] E. B. Moulin, *Spontaneous Fluctuations of Voltage* (Oxford University Press, London, 1938).

Substituting (32) into (31) obtains

$$\tau = \epsilon\lambda^2/3\mu kT. \quad (33)$$

It is felt that this derivation of τ gives a clear and convincing model of the Johnson noise. One can quickly obtain the same result by combining the well-known formula $\epsilon/m = \mu/\tau$ with the mean square velocity $\langle v^2\rangle$ from classical statistics,

$$\langle v^2\rangle = (\lambda/\tau)^2 = 3kT/m. \quad (34)$$

Assuming that N is known, the number of electrons ν which contribute to a pulse of a given height can be determined from (27) as

$$\nu = (N\tau' V^2/\tau\langle V^2\rangle)^{\frac{1}{2}}, \quad (35)$$

since τ' can be obtained from the experimental curve of (23) and τ can be obtained from (33). It is not surprising to note that if there were no limitations of amplifier resolving power, and (5) were substituted for (4), then it follows by (35) that

$$\langle \nu^2\rangle = N. \quad (36)$$

Furthermore, assuming still that $\tau' \geqslant \tau$, it follows by (32) that

$$\langle V^2\rangle/R = kT/\tau, \quad (37)$$

which has been derived by Bakker and Heller[3] using Fermi-Dirac statistics. By (37), the mean energy per pulse $\bar{E}$ equals

$$\bar{E} = kT, \quad (38)$$

and as the mean number of electrons involved in the pulse equals $(N)^{\frac{1}{2}}$ by (36), the energy contribution per electron E_0 is given by

$$E_0 = kT/(N)^{\frac{1}{2}} \quad (39)$$

which, in this model, represents the potential drop for an electron falling across a Lorentz mean-free path.

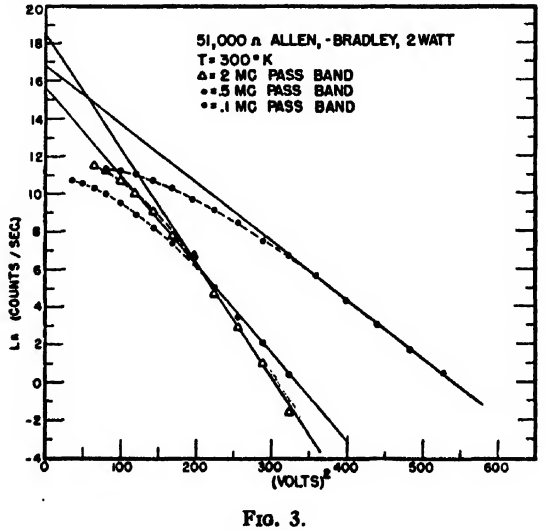


FIG. 3.

MEASUREMENTS

The measurements were taken with the following equipment all of which was made by the Instrument Department of the Oak Ridge National Laboratory: (1) A-1 linear preamplifier, (2) A-1 linear amplifier, (3) 64 scaler, Model Q-762, (4) 1024 scaler, Model Q-834. The A-1 linear preamplifier and amplifier have been described by Jordan and Bell.[12]

Pulse-height measurements taken with different types of commercial grid resistors are shown in Figs. 1–3; and these are the experimental curves which correspond to (23). Curve 5 (Fig. 1) with the input grid grounded gives the shot noise contribution. It can be seen from Figs. 1 and 2 that with the 2-mc pass band of the A-1 amplifier the resolution time of the system τ' extrapolates to $\sim 2.5 \cdot 10^{-8}$ sec. This is the time required by the pulse-height discriminator to receive enough energy to trigger, and consequently it is not inconsistent that it is less than the rise time of the amplifier, i.e., ~ 0.15 μsec. In measuring the random counting rate above $\sim 10^3$ counts/sec. a correction for the recovery time T was made using the equation[13]

$$n_1 = n(1-nT), \quad nT \ll 1,$$

where n_1 is the observed rate, and n is the true rate. The extrapolation of the curves in Figs. 1–3 is justified on the basis of this correction.

In order to predict the curves of Figs. 1–3 from Eq. (23) it is only necessary to have the parameters $\langle V^2 \rangle$ and τ' evaluated. $\langle V^2 \rangle$ can be determined by solving Nyquist's equation for a parallel RC circuit where C is the grid input capacitance, while τ' can be estimated from the design of the equipment.

[12] W. H. Jordan and P. R. Bell, Rev. Sci. Inst. 18, 703 (1947).

[13] See W. C. Elmore and M. Sands, *Electronics* (McGraw-Hill Book Company, Inc., New York, 1949), p. 207.

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to many people who have made valuable suggestions, and with whom he has had stimulating discussions on the subject. Special thanks are indeed due to Mr. P. R. Bell of ORNL and Mr. Keith Henry of NEPA.

Note on the Use of Reiner's Equation for Recognition of Pseudoplastics from Their Flow Curves

VERA OSMAN KRAKAUER
Interchemical Corporation, New York, New York

(Received March 17, 1950)

A method is described which has been found useful to determine by application of the Reiner equation whether certain flow curves obtained with a concentric cylinder rotational viscometer indicate pseudoplastic or Bingham body types of flow.

THERE has been an increasing use of concentric cylinder rotational viscometers suitable for obtaining complete rate of shear-shearing stress diagrams or flow curves, such as the one described by Green[1] from whose characteristics the rheological behavior of the measured sample may be determined. The several commercial viscometers of this type recently placed on the market are being used in diverse industrial fields, with the likelihood of encountering all types of rheological flow.

There is usually no difficulty in recognizing Newtonian flow from its curve of a straight line passing through the origin (Fig. 1A). Thixotropic plastics can be distinguished by the characteristic hysteresis loops formed between the slope of the up curve and the straight line of the down curve (Fig. 1B), particularly if measured to several top rates of shear,[2] and by the intercept on the shearing stress axis. However, non-thixotropic Bingham bodies, which have an intercept on the shearing stress axis, but then increase linearly with rate of shear (Fig. 1C), and pseudoplastics, whose flow curves show a less than linear shearing stress increase with increasing rate of shear, and which appear to start at the origin and to approach an asymptote (Fig. 1D) sometimes present a difficulty.

For pseudoplastics, the rheologist may be able to fit a straight line through the top points of the flow curve which will intersect the torque axis, with the lower points falling off, and obtain a seeming resemblance to Bingham-body flow curves where the lower points fall off due to plug flow caused by only partial shearing throughout the clearance, due to the yield value. In addition, over the practical range the apparent viscosity, which some investigators calculate as shearing stress divided by rate of shear at each point will continually decrease with increased shear for both types of flow. Therefore, at times, difficulties of recognition will arise. It is important to avoid erroneous conclusions and recognize the true nature of the sample in question that no misleading yield values be calculated for pseudo-

[1] H. Green, Ind. Eng. Chem. (Anal. Ed.) 14, 576 (1942).
[2] H. Green and R. N. Weltmann, Ind. Eng. Chem. (Anal. Ed.) 15, 201 (1943).

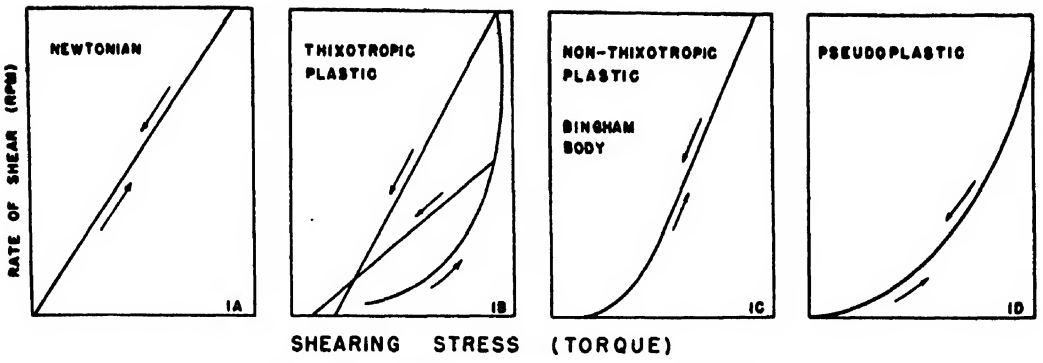


FIG. 1. Schematic flow curves for four types of flow.

plastics, which have none, or that no constant plastic viscosity be calculated for them.

Green points out this difficulty,[3] and suggests recourse to a microscopic method described previously,[4, 5] to help decide which type of flow is being encountered. This note describes a simpler, rapid method, that works well in the majority of cases, and requires merely an examination of the flow curve and a simple calculation.

Reiner and Rivlin, in the development of their equation of flow in concentric cylinder rotational viscometers point out[6] that for materials having a yield value, flow will not start immediately, but will only commence at a torque T_0 where,

$$T_0 = 2\pi r_b^2 hf$$

and completely laminar flow, which is never reached at finite flow rates in a capillary viscometer,[7] will commence in a rotational viscometer at a torque T_2 where

$$T_2 = 2\pi r_c^2 hf,$$

where r_c = radius of the cup (outer cylinder) in cm, r_b = radius of the bob (inner cylinder) in cm, f = yield value in dynes cm, h = immersed height of bob in cm. At T_2, for Bingham bodies the plug flow will be overcome and the straight-line portion of the curve, from whose slope the plastic viscosity, and from whose intercept (T_1) the yield value is calculated, will begin. This means that the individual points above torque T_2 should now fall directly on the straight line drawn through the upper points. Within limits of experimental error, this condition is realized for almost all of the collection of the non-thixotropic Bingham bodies measured at this laboratory over recent years. The measured points begin to fall on the line, at, or slightly above the calculated T_2 (Fig. 2). Thixotropic samples occasionally show larger deviations, possibly due to additional breakdown on the down curve.

However, when a straight line is fitted to the curve obtained for pseudoplastics to a certain top rate of shear, and T_2 is calculated for the yield value obtained from the intercept on the torque axis, it is seen that the individual points do not fall on the line until a much higher torque is reached than that attributable to plug flow if there were a yield value, often as much as 3 and 4 times the calculated value. Particularly when the measurements are continued to higher top rates of shear, it can be seen that the respective straight lines drawn will give materially higher intercepts with increasingly larger deviations from the respective calculated T_2 (Fig. 3), while for non-thixotropic Bingham bodies, measurements to successively higher rates of shear will

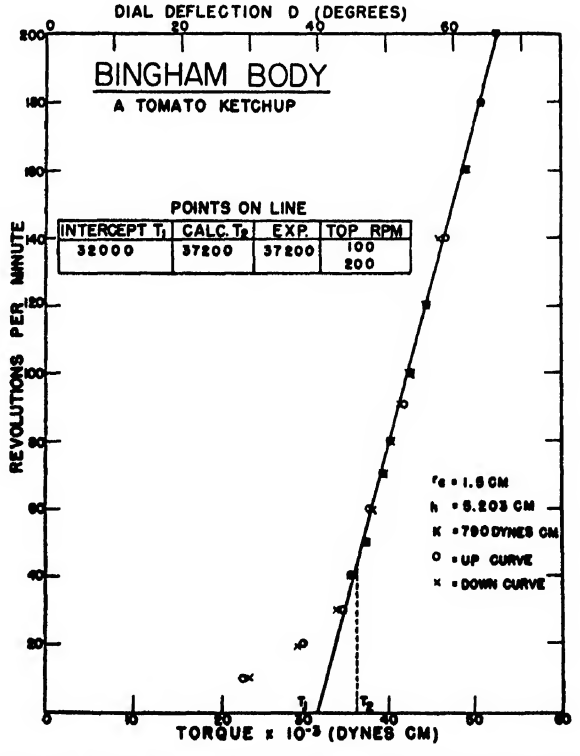


FIG. 2. Experimental flow curve obtained on rotational viscometer for a Bingham body, and calculations.

[3] H. Green, *Industrial Rheology and Rheological Structure* (John Wiley and Sons, Inc., New York, 1949).
[4] H. Green, Proc. A.S.T.M. 20, Part II, 451 (1920).
[5] H. Green and G. S. Haslam, Ind. Eng. Chem. 17, 7, 726 (1925).
[6] M. Reiner and R. Rivlin, Kolloid Zeits. 43, 1 (1927); J. Rheology 1, 5 (1929).
[7] E. Buckingham, Proc. A.S.T.M. 21, 1154 (1921).

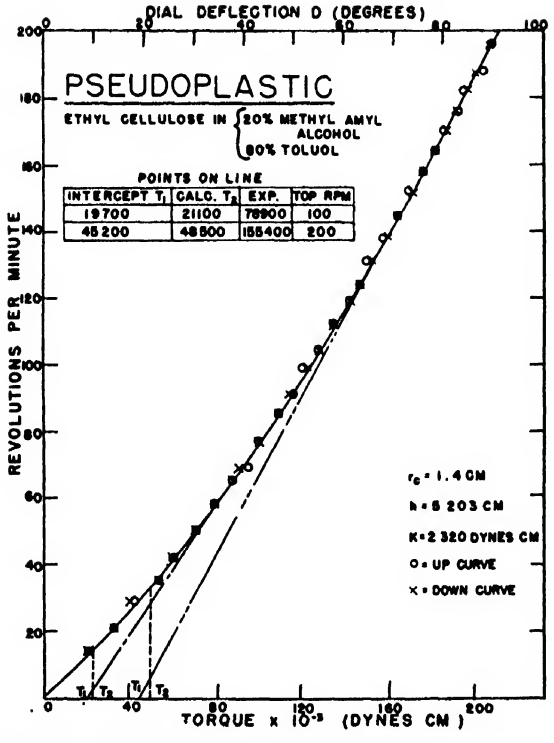


FIG. 3. Experimental flow curve obtained on rotational viscometer for a pseudoplastic, and calculations.

not affect either the yield value or T_2. Simple inspection of the flow curve and the number of points off the line will not usually suffice, since, depending upon the instrumental dimensions, the yield value, and the particular r.p.m. steps chosen, the number of points with too low a torque may vary for a particular Bingham body, while, depending upon the top r.p.m. chosen for the measurement of a pseudoplastic, and the number of points obtained, the amount of curvature may appear to vary. Therefore, it would be preferable to reach sufficiently high rates of shear and to calculate T_2 from the equation. When the flow curves for the McMichael type viscometer are simply plotted directly from the viscometer data as r.p.m. *vs.* angular deflection, changing the deflection values to torque can be avoided by dividing through by the torsion constant (K) of the spring, wire, or other torsion member used,

$$\frac{2\pi r_c^2 hf}{K} = D_2;$$

or, from the equation $f = T_1 C$, which is derived from the Reiner concept, where

$T_1 = D_1 K$,

D_1 = intercept on deflection axis,

$$C = \frac{(1/r_b)^2 - (1/r_c)^2}{4\pi h} \Big/ \ln \frac{r_c}{r_b} = \text{instrumental constant for yield value calculation;}$$

or one can substitute and altogether avoid calculation of f (which would be meaningless for pseudoplastics), using

$$2\pi r_c^2 hCD_1 = D_2,$$

where D_2 is the deflection at which laminar flow ensues and the point should fall on the line. Figures 2 and 3 were plotted from data obtained with a Precision-Interchemical viscometer at increasing and decreasing r.p.m. The coincidence of the "up" and "down" points indicates that no thixotropic breakdown took place. Both graphs could have remained in terms of deflection as indicated along the top of their abscissas. For the Bingham body (Fig. 2), the data for the lines at both top r.p.m. would have been $D_1 = 40.5$, $D_2 = 47$ and D actually on line also 47. For the pseudoplastic curve (Fig. 3), for the 100 top r.p.m. line $D_1 = 8.5$, $D_2 = 9.1$ and D actually on line $= 34$, and for the 200 top r.p.m. line $D_1 = 19.5$, $D_2 = 20.9$ and D on line $= 67$. Similarly, for the concentric cylinder modified Stormer type viscometer often used for non-thixotropic samples, whose flow curves are plotted as r.p.m. *vs.* weight where

$T_1 = (W_1 gS/G)$ and W_1 = intercept on weight axis
g = acceleration due to gravity
S = radius of drum to which weight is attached
G = gear ratio between drum and shaft of bob,

one can simply use

$$2\pi r_c^2 hCW_1 = W_2,$$

where W_2 is the weight after which for Bingham bodies the points should fall on the line.

Thus a convenient, rapid method is available to check when doubt exists as to whether a certain viscometric flow curve indicates a pseudoplastic or a Bingham body type of flow. When true yield value is present, the lower points will have too low a torque due to plug flow, but will fall on the straight line at or slightly after the torque at which completely laminar flow ensues, according to the Reiner equation. However, for pseudoplastics, where the intercept is simply a result of fitting a straight line to the upper portion of the curve resulting from the shear-produced reduction of torque, and where both the line and its intercept have no physical significance, the points will usually not fall on the line at or near the calculated value of T_2.

An X-Ray Study of Thermally Induced Stresses in Microconstituents of Aluminum-Silicon Alloys*

JOHN P. NIELSEN† AND WALTER R. HIBBARD, JR.‡

(Received March 13, 1950)

The changing width of high angle Debye x-ray diffraction lines from eutectiferous aluminum-silicon alloys during thermal cycling as measured by a recording Geiger counter spectrometer is interpreted in terms of stresses resulting from the differences in thermal expansion of the microconstituents.

IN a solid alloy containing a dispersed second phase, stresses theoretically should develop in the matrix during cooling from some stress-free state due to the difference between the thermal expansion coefficients of the two phases. If the matrix has the larger coefficient, the stress development process resembles that in shrink or interference fitting except that the components are microconstituents. László[1,2] developed the stress analysis techniques for a variety of such stresses, as they might occur in different ideal microstructures, labeling them "tessellated stresses." Metallographic evidence of the microstructural stresses reported by Boas and Honeycombe[3] consists of slip lines, cracks, and distortion produced by several heating and cooling cycles of certain alloys. The present work is a x-ray diffraction study of such stresses in aluminum-silicon alloys.

Eutectic and certain hypereutectic alloys of this system contain essentially silicon rich particles imbedded in an aluminum rich matrix below the eutectic temperature. Since the thermal expansion coefficient of aluminum is about seven times greater than that of silicon, the aluminum matrix tends to contract against the silicon particles on cooling these alloys, developing a pressure both in the matrix and in the silicon. If, as a first approximation, the silicon particles are considered as uniformly dispersed spheroids, the stress analysis of the matrix becomes that of the pressure on the inner wall of a thick spherical shell, producing radial compression and tangential tension, which depend upon the elastic moduli of the constituents and the dimensions of the shell. Based on physical constants from the 1948 *Metals Handbook* and the equations of László[1,2] and Love,[4] the calculated thermally induced stresses for certain of these alloys are given in Table I. These data indicate that the stresses may reach significant magnitudes and in real alloys where the silicon particles are far from spherical, the stresses at local points should readily exceed the yield strength of the matrix. It is to be noted that the difference between the radial and the tangential stresses at a point acts to produce plastic deformation, with the radial pressure considered as negative in sign.

EXPERIMENTAL PROCEDURE AND RESULTS

The measurement of the width of Debye x-ray diffraction lines appeared to be the best method for obtaining direct evidence of the thermally induced stresses in the aluminum matrix, since both tensile and compressive stresses would tend to shift the line slightly in both directions resulting in broadening. Aluminum-silicon alloys (12, 20, and 30 percent Si) were modified with NaF, chill cast, hot forged to further disperse the silicon particles, and annealed 20 hours at 500°C. A calibrated goniometer and a Geiger tube from a Philips x-ray spectrometer were assembled around a Seemann demountable-target-type x-ray tube so as to cover Bragg angles from 40° to 83°. By using a selected beam from a line focus source[5] fine slits of 0.005 in. at the source and at the Geiger tube permitted sufficient intensity for the delineation of line breadths in the back-reflection region. In order to take full advantage of the x-ray focusing principle,[6] the specimen was mounted so that the effective diffraction surface could be experimentally adjusted tangent to the focusing circle. To further enhance resolution, the scanning speed of the Philips goniometer was reduced to 0.05° Bragg per minute. Relative x-ray intensities were recorded with a chart speed such that one degree Bragg required 40 in. of chart. Line breadth measurements were made at several temperatures on cooling from 400°C to room temperature using the 420 diffraction line of the aluminum matrix and Cu$K\alpha$-radiation. The temperatures were maintained by an electrical resistance furnace which fitted over the specimen holder and contained narrow slits for the exit and entry of the beam. The temperature

* From a dissertation presented by John P. Nielsen to the Graduate School of Yale University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

† Associate Professor of Metal Science, New York University, New York, New York.

‡ Assistant Professor of Metallurgy, Yale University, New Haven, Connecticut.

[1] F. László, J. Iron and Steel Inst., Part I, No. 1, 173P (1943); Part II, No. 2, 137 (1943); Part III, No. 2, 183P (1944); Part IV, No. 2, 207P (1945); Part V, 164, 5 (1950).

[2] F. R. N. Nabarro, László's Papers on Tessellated Stresses, Symposium on Internal Stresses in Metals and Alloys, Inst. of Metals, London (1948).

[3] W. Boas and R. W. K. Honeycombe, Proc. Roy. Soc. 186, 57 (1946); 188, 427 (1947). J. Inst. Metals (January, 1947). Council for Sci. and Ind. Res., Melbourne, Australia, Serial No. A. 151, Physical Metallurgy Report No. 4 (May 6, 1947). Nature, 153, 494 (1944); 154, 388 (1944).

[4] A. E. H. Love, *Mathematical Theory of Elasticity* (Cambridge University Press, London, 1920); third edition, p. 140.

[5] N. C. Baese, Rev. Sci. Inst. 8, 258 (1937).

[6] J. C. M. Brentano, Proc. Phys. Soc. 37, 184 (1925).

TABLE I. Calculated microstructure shrink-fit stresses developed on cooling from 300°C to different temperatures in aluminum shell-silicon core compound sphere. Stresses in 1000 p.s.i.

Wt. percent Si	To: 200°C	Tangential stress (tension) Inner surface 100°C	20°C	200°C	Outer surface 100°C	20°C	Radial stress (compression) Inner surface[a] 200°C	100°C	20°C
12	14	29	41	4	8	11	20	42	60
20	14	29	41	4	8	11	20	42	60
30	17	35	49	8	17	24	17	35	49
40	19	39	55	12	25	35	13	27	38

[a] Radial stress at outer surface of aluminum shell vanishes.

was measured by a thermocouple placed in a hole drilled in each specimen and was controlled to within ±2°C. The deviation in line breadth produced by the 2°C variation was negligible. The results are tabulated in Table II.

TABLE II. Widths in radians, W_r, at half-maximum for the 420 Cu$K\alpha_1$-aluminum line from aluminum-silicon alloys at various temperatures (multiply by 10^{-3}).

0 percent Si		12 percent		20 percent		30 percent	
T°C	W_r	T°C	W_r	T°C	W_r	T°C	W_r
260	3.5	420	3.5	307	1.6	313	3.5
22	2.6	320	3.5	272	2.5	260	3.2
		262	3.5	214	3.2	206	5.1
		211	3.0	182	3.5	153	5.1
		142	3.3	144	3.2	96	5.4
		77	4.2	97	3.3	24	5.6
		12	4.0	31	3.7	360	4.3
				22	3.9	22	5.2

A metallographic confirmation of the stresses was attempted¶ similar to the Boas and Honeycombe experiments. It was found that polished specimens of these alloys would not develop slip lines[7] on thermal cycling, but neither could slip lines be developed by mechanical straining (squeezing in a vise), so that it does not appear possible to obtain metallographic support for the x-ray data.

EVALUATION OF RESULTS

As expected, cooling the aluminum (0 percent Si) specimen from 260°C to room temperature produced no line broadening, in fact, the high resolution even revealed a line sharpening. However, all alloys containing added silicon showed line broadening on cooling, as predicted by the stress analysis. No pronounced asymmetry was found in the shape of the lines, indicating no measureable difference in the magnitude of the tensile and compressive components. On reheating the 30 percent silicon specimen, it was found that the line width decreased somewhat, lending support to the thesis that elastic stresses were being detected. The gradient nature of the stresses together with the fact that the silicon particles were not spherical precluded any quantitative correlation with the calculated stresses. Furthermore, the x-ray measurements involved only the region at the surface of the specimen. Regarding the solid solubility of silicon in the aluminum phase, calculations indicate that the 0.05 percent to 0.48 percent solubility limits between room temperature and 400°C would disturb the x-ray lines less than the probable error of measurement. However, the gradient nature of the stresses and any cold working resulting from stresses exceeding the yield point of the matrix might have added to the line broadening on cooling.

¶ Experiment performed by W. H. Hawley, Jr., graduate student at Hammond Laboratory, Yale University.

[7] A "rumpled" surface indicating strain was visible to the naked eye, but, under the microscope, fine dark lines indicative of slip were not visible. For a similar phenomenon see Guy, Barrett, and Mehl, Trans. A.I.M.E. 175, 216 (1948).

SUMMARY

Thermally induced microstructure shrink-fit stresses produced in the aluminum matrix around silicon particles in duplex aluminum-silicon alloys are analyzed on the basis of elastic theory, and x-ray diffraction experimental evidence is presented to confirm their presence.

ACKNOWLEDGMENT

Acknowledgment is due the North American Philips Company, New York, New York for their generous loan of the Geiger counter spectrometer. The personal efforts of Dr. O. S. Duffendack, Director of Research, Dr. W. Parrish, and Mr. C. J. Woods, of that company in arranging for the loan are particularly appreciated. Dr. Kent R. Van Horn of the Aluminum Company of America, Cleveland, Ohio, arranged for the supply of specimen material. Professor L. W. McKeehan, Director of Sloane Physics Laboratory, Yale University and Dr. C. H. Mathewson Department of Metallurgy, Yale University kindly read the manuscript and offered helpful suggestions.

X-Ray Diffraction by Elastically Deformed Crystals*

J. E. White
*Acoustics Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts***
(Received March 20, 1950)

Published articles establish the fact that vibration or other inhomogeneous strain greatly increases the ability of a crystal to diffract x-rays. The present article gives experimentally determined relations between integrated intensity and strain for the rather simple case of static bending. All measurements were made on quartz plates, but a wide range of geometrical and crystalline parameters was covered nevertheless. In some cases, the integrated intensity for extreme bending was more than 20 times that from the unbent crystal. An approximate theory is developed which agrees well enough with the measurements to serve as a means of computing the effects of other types of strain or of other combinations of crystalline parameters. A qualitative study of vibrating crystalline plates is also described briefly.

NOMENCLATURE

e	electronic charge
m	electron mass
c	velocity of light
N	number of unit cells per unit volume
λ	x-ray wave-length
F	crystalline structure factor
θ	Bragg angle
ϕ	variable angle of incidence
l	distance traversed in crystal
R	radius of bending
$2s$	angular width for perfect reflection
w	"average" width
ϵ	extinction coefficient
ϵ_{ave}	average extinction coefficient
μ	linear absorption coefficient
A_0	integrated intensity for perfect crystal
B, L	integrated intensities
t	thickness of crystal
Θ	angle between crystalline plane and face of crystal
I	beam intensity
P_0	beam power
ω	angular velocity of crystal
E	energy diffracted in one traversal

INTRODUCTION

THE influence of elastic vibrations of a crystal on the diffraction of x-rays by the crystal has interested many investigators during the last twenty years.[1] One fact clearly established by their work is that Laue spots formed when an x-ray beam traverses a crystal vibrating in some resonance condition are much more intense than they are when the crystal is at rest. A similar increase for Bragg reflection off the surface was found. Statically bent quartz was also shown to exhibit this effect. The increased intensity was generally attributed to a reduction in extinction, either primary or secondary. A theory[2] treating diffraction by a crystal whose planar spacing varies sinusoidally in space and time was published, but extinction effects were not considered, and as a result, the integrated intensity of any reflection came out to be independent of vibration.

The principal objective of the work being described here is the determination of a quantitative relationship between increased diffracting power and degree of deformation. Since high frequency vibration patterns may be quite complex, static bending was chosen as a type of strain which can be realized experimentally and described easily. Quartz was chosen for study because of its excellent elastic properties and the high degree of crystalline perfection obtainable. The experimental procedure and results will follow the discussion of the approximate theory, since the terminology and basic objectives can best be presented in connection with the theory. The expressions derived provide a means of generalizing the results to other crystals and other types of strain, including some forms of vibration. No measurements were made on vibrating crystals, although the qualitative experiments performed provide further evidence that it is the local curvature of the crystal which controls the increase in diffracting power.

AN APPROXIMATE THEORY

The mathematical description of x-ray diffraction by a distorted crystal as developed here relies upon the Darwin[3] treatment of x-ray diffraction by a perfect crystal for fundamental constants required. He found that a plane monochromatic beam is perfectly reflected over a small angular range which can be expressed (for σ-polarization),

$$2s=(2e^2N\lambda^2F)/(mc^2\pi\sin 2\theta). \tag{1}$$

Outside this range, the reflected amplitude drops sharply. Even so, the area under the "tails" is one-third the area under the flat portion of the curve. The area under this curve is customarily expressed in units applicable to an experiment in which the crystal is rotated at a uniform angular velocity ω and the total energy E is measured as the region of reflection is traversed once. Expressed as a ratio to the incident power P_0,

$$\frac{E\omega}{P_0}=A_0=\frac{8N\lambda^2Fe^2}{3\pi\sin 2\theta mc^2}\frac{1+\cos 2\theta}{2}. \tag{2}$$

* Extract from a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology.

** Now at the Magnolia Petroleum Company, Field Research Laboratories, Dallas, Texas.

[1] W. G. Cady, *Piezoelectricity* (McGraw-Hill Book Company, Inc., New York, 1946).

[2] J. Weigle and K. Bleuler, Helv. Phys. Acta 15, 445 (1942).

[3] C. G. Darwin, Phil. Mag. 27, 315 and 675 (1914).

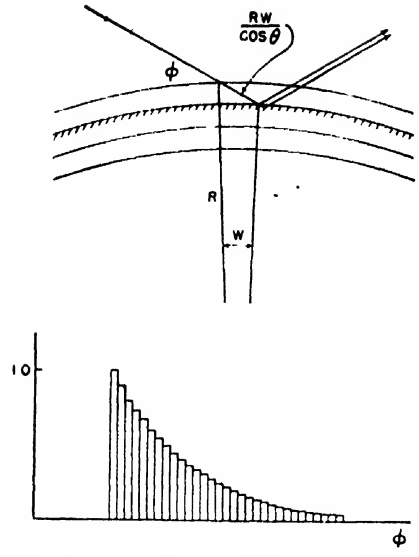


FIG. 1. Geometry for approximate theory (above) and resulting rocking curve (below).

This is the integrated intensity for a perfect crystal when the incident beam is unpolarized.

Darwin also found that the transmitted beam is rapidly "extinguished" within this angular range. The reduction in amplitude as the beam penetrates the crystal along a path l can be expressed as an exponential decay,

$$I_l = I_0 e^{-\epsilon l},$$

where

$$\epsilon = (2\pi \sin 2\theta)[s^2 - (\phi-\theta)^2]^{\frac{1}{2}}/\lambda. \qquad (3)$$

Within the region of perfect reflection, the extinction coefficient ϵ varies, being a maximum at the middle and zero at the edges. Outside the region of perfect reflection, ϵ is imaginary and therefore does not represent attenuation. The values of ϵ are such that reflection by a perfect crystal is practically a surface phenomenon.

The essential feature of the theory described here is the assumption that a distorted crystal behaves locally like a perfect crystal, the strains becoming apparent only when viewed at points many wave-lengths apart. It is developed for a crystal bent into a section of a cylinder of radius R. An expression for the area under the rocking curve as a function of bending is our objective. Before proceeding, let us make two simplifying approximations regarding the behavior of a perfect crystal. The first is that perfect reflection occurs over a range w such that the area A_0 is the same as that given by the Darwin treatment. This makes $w = 2.66s$ as compared with $2s$. The second assumption is that the extinction coefficient ϵ, which governs the depth of penetration of the beam before it is reflected, has the constant value ϵ_{ave} over the whole range w and is zero elsewhere. ϵ_{ave} is the linear average of Darwin's ϵ over the angular range w, and it is given by

$$\epsilon_{Av} = \frac{3\pi^2 \sin 2\theta s}{8\lambda} = \frac{3\pi e^2 N \lambda F}{8mc^2}. \qquad (4)$$

We are now ready to state how the integrated intensity from a bent crystal should depend on the radius of bending. As a bent crystal (Fig. 1) is rotated counter-

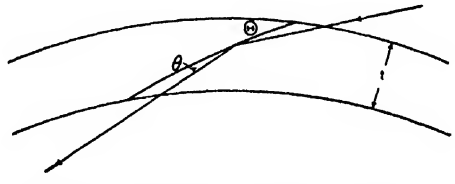


FIG. 3. Geometry for plane not parallel to surface of crystal.

clockwise, the conditions for reflection of the elementary pencil of x-rays shown will become satisfied for the region near the surface. Throughout a range w, reflection is perfect, and a contribution of A_0 is made to the integrated intensity. The beam penetrates only a very short distance controlled by ϵ_{ave}, and the rest of the crystal is shielded. As the angle proceeds beyond w, however, the beam is no longer reflected at the surface, but can penetrate with attenuation due only to ordinary absorption. After penetrating a distance $Rw/\cos\theta$, it reaches a region which has been bent through w with respect to the surface, counteracting the rotation w already given to the whole crystal, so conditions are right for reflection. (Since the whole angular range for reflection is very small, the angle can be considered constant at the Bragg angle θ for purposes of geometry.) During the next w radians, A_0 is contributed by the shaded area in Fig. 1, but the beam is reduced by $\exp[-(2\mu Rw/\cos\theta)]$ in twice traversing the surface layer, where μ is the linear absorption coefficient. In this fashion, the crystal is considered made up of discrete layers, each reflecting like a perfect crystal. This proceeds until the back side of the crystal is properly oriented, or until the total distance is $t/\sin\theta$, t being the thickness of the crystal. This process can be stated as a

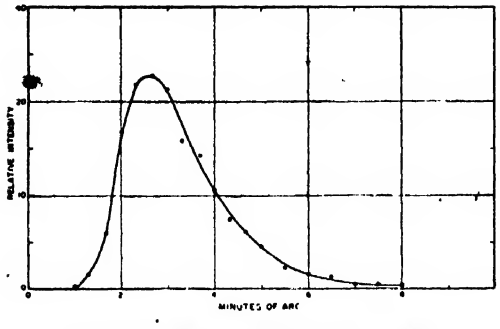


FIG. 2. Rocking curve for bent crystal.

summation,

$$B_0 = A_0 \sum_{n=0}^{\frac{t\cos\theta}{Rw\sin\theta}} \exp[-n2\mu Rw/\cos\theta],$$

$$B_0 = A_0[1-\exp(-2\mu t/\sin\theta)]/[1-\exp(-2\mu Rw/\cos\theta)].$$

It is apparent that as bending is increased, the layers described above become thinner, whereas the distance required for almost complete reflection remains fixed. In Fig. 1, the shaded area becomes deeper than a layer as bending becomes extreme. This "thin crystal" effect reduces the contribution of each layer and causes the curve of integrated intensity *versus* bending to flatten off. A statement of this effect was derived in the following way. In a thick perfect crystal, the transmitted beam is reduced exponentially, as derived by Darwin. Let us state that this is true for our crystal, and that ϵ_{ave} is the proper exponent. Thus $dI/I = -\epsilon_{ave}dl$. Let us further state that the rapid loss of energy from the transmitted

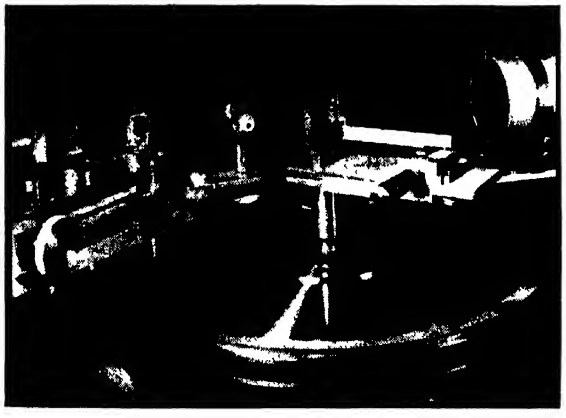


FIG. 4. Apparatus including monochromator, bent crystal, and counter.

beam is due to its transfer to the reflected beam, so that the rate of increase of the reflected beam equals the rate of loss from the transmitted beam.

$$dI_R = -dI = I_0\exp(-\epsilon_{ave}l)\epsilon_{ave}dl,$$

$$I_R = I_0\int_0^{Rw/\cos\theta} \exp(-\epsilon_{ave}l)\epsilon_{ave}dl$$

$$= I_0[1-\exp(-RW\epsilon_{ave}/\cos\theta)].$$

I_R/I_0 is the factor by which reflection is reduced when a crystal becomes thin. Hence, a complete expression for integrated intensity from a bent crystal is

$$B = A_0[1-\exp(-2\mu t/\sin\theta)] \times[1-\exp(-\epsilon_{ave}wR/\cos\theta)]/[1-\exp(-2\mu wR/\cos\theta)]. \quad (5)$$

The step-wise procedure described above would lead to a rocking curve as sketched at the bottom of Fig. 1. A

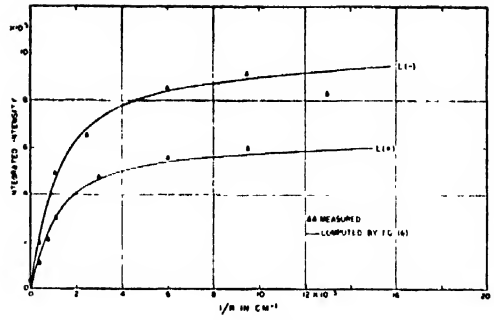


FIG. 5. Integrated intensity for 20.0 plane at 38°13′ with crystal face using Mo$K\alpha$.

highly monochromatic beam extremely limited in divergence and area gave the experimental curve in Fig. 2, which bears a strong resemblence to the shape suggested. The rocking curve obtained for no bending was narrower and symmetrical, and bending in the opposite direction gave correspondingly reversed asymmetry.

Although symmetrical Bragg reflection probably represents the most frequent usage of a crystal, it is nevertheless a special case, and it seemed desirable to work out an expression similar to the one above for diffraction by crystalline planes making an arbitrary angle with the surface of the crystal. The treatment is limited to crystalline planes which are parallel to the axis of bending. The geometry shown in Fig. 3 brings out the fact that an interchange of source and receiver changes the angle between the incident beam and the surface of the crystal. It turns out that the two cases lead to different expressions for integrated intensity, a result which may at first seem surprising. If the diffracted beam emerges on the same side of the crystal as the incident beam, corresponding to $\Theta<\theta$, it is referred to here as a Bragg reflected and designated by B. Similarly, if the beam traverses the crystal, corresponding to $\Theta>\theta$, it is called a Laue reflected and designated by L. In each case, plus and minus signs are added to differentiate the two possible angles of incidence. The steps leading to the expressions given below will not be in-

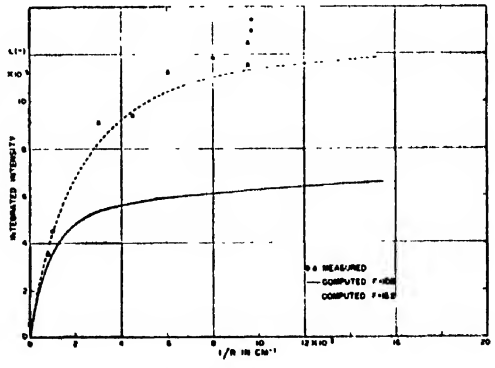


FIG. 6. Integrated intensity for 10.0 plane at 38°13′ with crystal face, using Mo$K\alpha$-radiation.

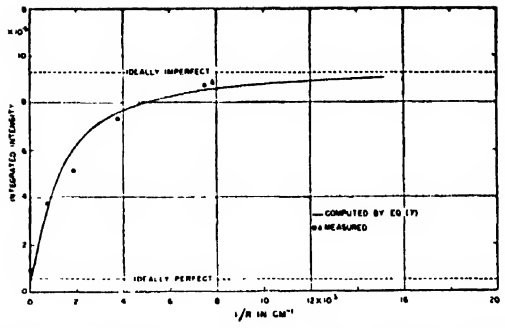


FIG. 7. Integrated intensity for 20.2 plane parallel to crystal face, using Mo$K\alpha$-radiation.

cluded, since the more involved geometry makes the derivation rather lengthy without adding anything to the viewpoint used in the simpler derivation.

$$L(\pm)=A_0\exp[-\mu t/\sin(\Theta\mp\theta)]\times\{1-\exp[\pm\mu t(\sin(\Theta+\theta)-\sin(\Theta-\theta))/\sin(\Theta+\theta)\sin(\Theta-\theta)]\}\times\{1-\exp[-\epsilon_{ave}wR/\cos\theta\cos\Theta(1+\sin^2\Theta)]\}/\{1-\exp[\pm\mu wR(\sin(\Theta+\theta)-\sin(\Theta-\theta))/\sin(\Theta\mp\theta)\cos\theta\cos\Theta(1+\sin^2\Theta)]\}. \quad (6)$$

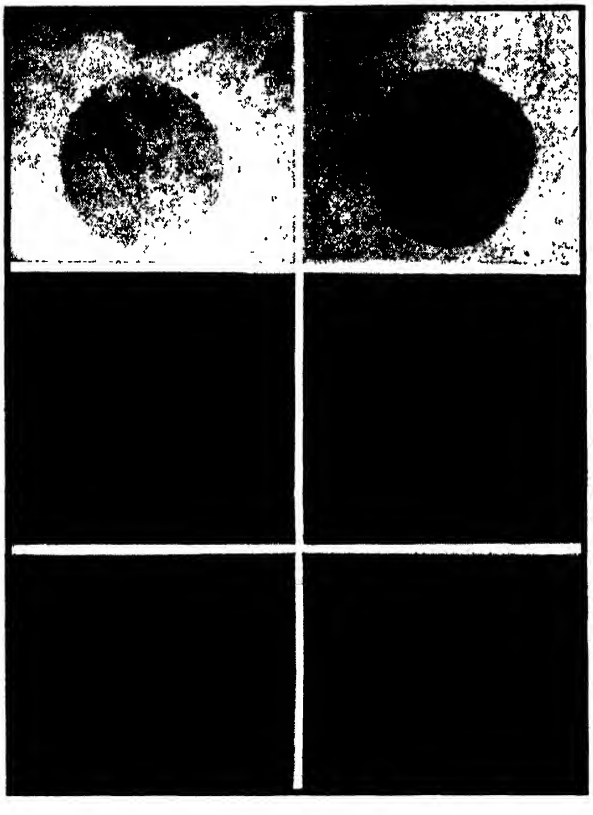


FIG. 8. Qualitative comparison of vibrating crystals (right) with the same crystals at rest (left).

$$B(\pm)=A_0\{1-\exp[-\mu t(\sin(\theta+\Theta)+\sin(\theta-\Theta))/\sin(\theta+\Theta)\sin(\theta-\Theta)]\}\{1-\exp[-\epsilon_{ave}wR/\cos\theta\cos\Theta(1+\sin^2\Theta)]\}/\{1-\exp[-\mu wR(\sin(\theta+\Theta)+\sin(\theta-\Theta))/\div\sin(\theta\pm\Theta)\cos\theta\cos\Theta(1+\sin^2\Theta)]\}. \quad (7)$$

In either case,

$$t\cos\theta\cos\Theta(1+\sin^2\Theta)/Rw\sin(\theta\pm\Theta)>5.$$

These are the expressions which have been compared with experiments on bent quartz. The assumptions made appear rather extreme. In particular, considering the contributions from different depths as a step-wise process is not justified unless there are many contributions. This means that the expressions are not applicable if the bending is too small, as indicated by the quantity following the expressions for integrated intensity. Setting this quantity to be greater than 5 rather than some other integer is arbitrary.

EXPERIMENTS WITH STATICALLY BENT QUARTZ

The apparatus for observing bent quartz is shown in Fig. 4. The beam strikes two calcite crystals set up as a double crystal spectrometer in the antiparallel position, thus achieving a wave-length spread in the final beam even narrower than the natural width of the α_1-line. A slit limits vertical divergence and beam area. The quartz slab is mounted on the axis of a precision turn-table in such a manner that motion of an ordinary micrometer at the end of a lever results primarily in a bending moment. Radius of bending computed from the lever position checked determinations by two other methods. The diffracted beam entered a Geiger tube which was mounted on a second turn-table geared to turn at twice the angular velocity of the first. Readings were taken at closely spaced angular positions by registering the number of counts in an accurately controlled minute by means of a scaling circuit. A plot of these readings, in arbitrary power units against angle in radians, constitutes a rocking curve. The power of the direct beam in the same units was measured by the Geiger tube and counter, although it was necessary to introduce a known absorption in the beam or effect a known reduction in tube output of a factor of 10 or so. The area under each rocking curve was divided by the incident power to obtain integrated intensity. Convex and concave bending were plotted together, and in agreement with the theory, have the same effect on integrated intensity. Measurements were made on two crystals with faces parallel to the 10.1 planes, and a third crystal with faces parallel to the 10.0 planes. In each case, the crystal was lapped to a thickness of about 0.05 cm and then etched for about two hours in 48 percent hydrofluoric acid. Fourteen curves were measured and computed, involving six crystallographic planes, some parallel to the crystal face and some not. Most of the curves were obtained with Mo$K\alpha_1$-radiation, the others with Cu$K\alpha$.

Agreement between computation and measurement was generally good, five of the curves being comparable with the two shown in Fig. 5. These curves illustrate the effect of interchanging source and receiver. Real discrepancies were found for all three curves involving the 10.0 plane, of which Fig. 6 is an example. The solid curve is based on the published structure factor.[4] The value of 10.8 for the 10.0 plane appears to be in error. A value of 16.3 brings computed curves into agreement with all three measured curves, and a computation based on the published structure yielded 15.2, suggesting that an error was made in the original calculation. The dashed curve is computed using 15.2 as the structure factor.

It is reasonable to suppose that the integrated intensity for any crystal lies between the value for perfection and the value resulting from such complete imperfection that extinction is of no consequence.[5] The former limit is given as Eq. (2). The integrated intensity for the ideally imperfect condition is

$$\frac{E\omega}{P_0}=\frac{N^2F^2\lambda^3e^4}{2\mu m^2c^4\sin 2\theta}\,\frac{(1+\cos^2 2\theta)}{2}. \tag{8}$$

These limits are shown as dashed lines in Fig. 7. The measured values almost cover the extreme range, and the computed curve is seen to approach a limit very near the ideally imperfect value.

EXPERIMENTS WITH VIBRATING CRYSTALS

Thin crystalline slabs of quartz and calcite were driven at frequencies of flexural resonance by means of a slender piezoelectric crystal cemented to the back faces of the slabs. The front faces were illuminated by a divergent beam of x-rays from a molybdenum target. The diffracted x-rays were observed on a film set parallel to the crystal and fairly near to the crystal. The crystal was rotated through an angular range such that each point had an opportunity to diffract the $K\alpha$-doublet. The film was mounted on an arm from a turn-table rotating at twice the angular velocity given the crystal. This procedure "scanned" the face of the crystal and gave an approximate picture. In each case, the crystals were etched before use.

Figure 8 shows three pairs of such pictures, illustrating the effect of vibration. The exposure times and photographic histories of both members of each pair are approximately the same, so the increased blackening observed is due to vibration. Other pictures of this kind pretty well established that radius of curvature is the parameter which controls this increased blackening. A more detailed description of the vibration studies is being submitted to the *Journal of the Acoustical Society of America*.

SUMMARY

The measurements given here show the increase in diffraction due to bending to be a gross effect, even a factor of 20 or more. It seems likely that this mechanism must contribute to the performance of bent crystal monochromators. The theory, although admittedly approximate, provides a means of predicting the effect of bending on various crystals whose crystal constants are known. The same procedure can be used to derive expressions for types of strain other than bending. The computed curve may also be used to compute a time average of diffracted energy when the radius is varying sinusoidally with time, as in flexural vibration.

ACKNOWLEDGMENT

The author wishes to express his appreciation for the suggestions and encouragement so generously given by Professor B. E. Warren during the course of this research and to thank Dr. W. Parrish of the Philips Laboratories for donating the quartz crystals used.

[4] P'ei-Hsiu Wei, Zeits. f. Krist. **92**, 355 (1935).
[5] W. A. Wooster and G. L. MacDonald, Acta Cryst. **1**, 49 (1948).

Some Effects of Nozzle Design on the Diffraction of Electrons by Gases

R. B. Harvey, F. A. Keidel, and S. H. Bauer
Department of Chemistry, Cornell University, Ithaca, New York
(Received March 20, 1950)

An evaluation has been made of some of the errors introduced into the electron diffraction of gases by failure of the sample to be localized, and by secondary and higher order scattering.

The degree to which poor nozzle design may contribute to these factors was investigated experimentally by measurements of the pressure in the camera and absorption measurements of the main electron beam when a small volume of gas was allowed to expand into the camera; and theoretically for the cumulative effects of higher order scattering on the final picture. This evidence was complemented by density measurements on photographic plates taken under similar conditions. The effect of using liquid air cooled surfaces around the nozzle was tested.

Of the nozzles tested, better "image to background" ratios were obtained with the "hydrodynamic" and "comparison" nozzles than with the "conventional" nozzle, but the gas must be admitted slowly enough so that less than about 10 percent of the main beam is appreciably scattered. The "image to background" ratios were always higher using adequate cooling on the condensing surfaces surrounding the nozzle. In the case of pressure and electron beam absorption measurements short exposures were advantageous, where secondary scattering was not directly considered, and high camera pumping speeds were always advantageous.

On the basis of these results some general design suggestions for electron diffraction cameras, and requirements for good efficiency in nozzles can be made. A nozzle design is proposed which would reduce materially the errors normally encountered in electron diffraction work.

INTRODUCTION

WHEN a stream of high velocity electrons is scattered by a gaseous sample, the intensity of the scattered radiation may be written:[1]

$$I_s = J_0\left[\frac{8\pi^2 me^2}{h^2}\right]^2 \frac{N}{R^2 s^4}[\text{atomic contribution} + \text{molecular contribution}], \quad (1)$$

where I_s can be defined in terms of the number of electrons crossing unit area per unit time at the focal plane of the diffracted ray; J_0 is the total intensity in the primary beam; N is the total number of molecules intercepted by the beam per unit beam area; $s=(4\pi/\lambda)\times\sin(\theta/2)$ where λ is the electron equivalent wavelength and θ is the angle between the primary and scattered rays; R is the distance from the sample to the focal plane; and $8\pi^2me^2/h^2=2/A_H$ where A_H is the "radius" of the hydrogen atom, and m, e, and h are universal constants. The "atomic contribution" factor is dependent only on the scattering functions of the individual atoms of the molecule, and so accounts for the structure independent contribution to the total scattered intensity, whereas the "molecular contribution" factor is in addition dependent on the interatomic distances in the molecule and so accounts for the structure dependent contribution to the scattered intensity.

The following are some of the assumptions made in the derivation of the above equation:

(1) The N molecules scatter independently with no mutual phase relations. This is satisfied for a gaseous sample.

(2) The sample is of infinitesimal thickness; diffraction occurs at one value of R only. This is practically never true for a gaseous sample.

(3) There is no secondary or multiple scattering (i.e., no scattered electrons are rescattered) in the sample region or in the space between the sample and the detector. This is generally not true for diffraction by a gaseous or a solid sample.

(4) The diffracted beam is perfectly focused at the plane of the detector. This is only approximately true.

A theoretical discussion of the errors which result when these assumptions are invalid appeared in a report[2] written for the Office of Naval Research.

The effects of various gas distributions about the point of gas injection were treated from the standpoint of image resolution, and it was found that indeed a loss of resolution was experienced, the loss becoming increasingly serious with large scattering angles, particularly for the high frequency terms. With regard to the use of a beam trap close to the nozzle to remove the main beam and to reduce the observed scattered intensity at small angles preferentially, so that the density range of the photograph may be reduced and excessive halation prevented, it was found that the positions of the inner maxima and minima of the pattern were not only subject to a similar loss of resolution but also to some lateral shifting of their positions. Experimentally it is also found that the use of a beam trap close to the nozzle leads to further difficulties insofar as edge scattering contributes appreciably to the general background.

A mathematical analysis of multiple scattering within the nozzle indicated that secondary scattering expressed as the ratio of [gain in background/loss in diffracted intensity] is dependent on the diffracting angle, although the fractional loss of any diffracted ray is probably independent of angle. The extent of multiple scattering was found to be independent of the sample distribution along the direction of the electron

[1] P. J. W. Debye, J. Chem. Phys. 9, 55 (1941).

[2] Bauer, Keidel, and Harvey, "An evaluation of quantitative procedures for the estimation of intensities of diffracted electrons," a report issued on work done under Contract No. N6 ori-213, Task Order I, NR 052-040. This report is available from the Library of Congress, Washington, D. C.

beam but to be proportional to the sample density at the point of diffraction.

The treatment of the last assumption, that pertaining to the effects of imperfect focusing at the plane of the detector, showed that although there was no appreciable shifting of the peaks or valleys of the intensity curves, there was a loss of resolution (the ratio of observed peak heights to the theoretical peak heights was less than 1), and this loss increased with increasing distance for any atomic pair. A slight additional loss in resolution which has angular dependence is also introduced by the use of a flat detector (e.g., a photographic plate) instead of one curved to fit the focal circle.

Although Eq. (1) has been checked quantitatively, within rather broad limits of experimental error,[3] the cumulative effects of all of the specific errors listed above must be carefully evaluated before one may determine precisely the structures of molecules which must be formulated in terms of more than two structural parameters.

The extent of invalidity of assumptions two and three is particularly influenced by nozzle design. It is from this viewpoint that the subject of nozzle design is discussed in the present paper. Some aspects of this study also appeared in the previously mentioned ONR report.[2]

NOZZLE EFFICIENCY DEFINITIONS

Experimentally, in order to obtain a gas diffraction pattern, a jet or stream of gas is allowed to intercept the electron beam. It is therefore apparent that the distance along which electrons are scattered cannot be smaller than that dimension of the gas exit which is parallel to the direction of the electron beam. Actually the gas never leaves the nozzle in a perfectly collimated beam but rather tends to diffuse throughout the camera. Thus, contrary to assumption (2), one obtains, in practice, a superposition of patterns originating at various distances R, rather than a single pattern originating only at R_0, the distance of the nozzle axis to the plane of the detector. Maximum resolution will be obtained only if the spread in R over which appreciable scattering takes place is as small as possible, i.e., when the function $N(R)$ is negligibly small except in the vicinity of R_0. However, a nozzle having a thin slit perpendicular to the electron beam will not be satisfactory in this respect because the issuing gas stream is then poorly collimated. In any case, an estimate of the amount of extraneous scattering could be obtained if the distribution of the gas in the camera could be plotted. It appeared reasonable that if the gas stream intersecting the electron beam directly over the nozzle gave rise to the principal diffraction pattern, then this might be called "useful diffraction," while the gas existing throughout the camera volume would give rise to many overlapping patterns, thus constituting a background, and this diffraction might be termed "harmful." One measure of the ratio of the "useful diffraction" to the total diffraction is then

$$E_p(t)=\frac{Q_{\text{nozzle}}}{Q_{\text{nozzle}}+Q_{\text{camera}}},$$

where Q_{nozzle} is the fraction of the total available sample that is in the electron beam directly over the nozzle, averaged up to time t; and Q_{camera} is the fraction of the total available sample that is in the rest of the electron beam (considered as a cylinder) averaged up to time t. One sees from reference to Fig. 1 that the ratio of the "useful diffraction" to the total diffraction is actually equal to

$$\frac{C_1+C_3}{C_1+C_2+C_3+C_4},$$

where C_n is the relative contribution of process n in Fig. 1, to the total intensity. Therefore a high efficiency $E_p(t)$ indicates that relatively little extension of the sample occurs beyond the dimensions of the nozzle and hence that the loss of resolution due to sample extension alone is small, regardless of how much resolution is lost as a result of multiple scattering (process 3).

We shall also discuss another measurement of efficiency:

$$E_{ab}(t)=\frac{C_1+C_3+C_4}{C_1+C_2+C_3+C_4},$$

which may be estimated from data on the absorption of the main electron beam. Values of $E_{ab}(t)$ should agree fairly well with corresponding values of $E_p(t)$ because process 4 contributes only slightly to the total scattered intensity. It is convenient to distinguish (2a), that part of process 2 taking place inside the chimneys surrounding the nozzle from (2b), that part taking place outside the chimneys. Then, if no liquid air cooling is used, we may write:

$$E_{ab}\text{ (no liquid air)}=\frac{C_1+C_3+C_4}{C_1+C_{2a}+C_{2b}+C_3+C_4},$$

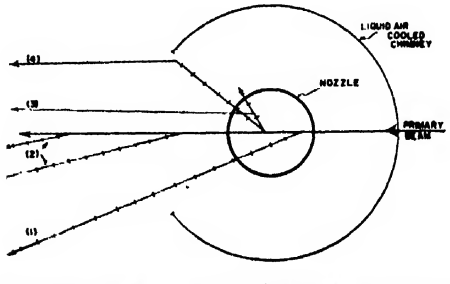


FIG. 1. Schematic representation of various scattering processes (1) Primary diffracted ray—within nozzle. (2) Primary diffracted ray—outside nozzle. (3) Secondary diffracted ray—within nozzle. (4) Secondary diffracted ray—outside nozzle.

[3] J. M. Hastings and S. H. Bauer, J. Chem. Phys. 18, 13 (1950).

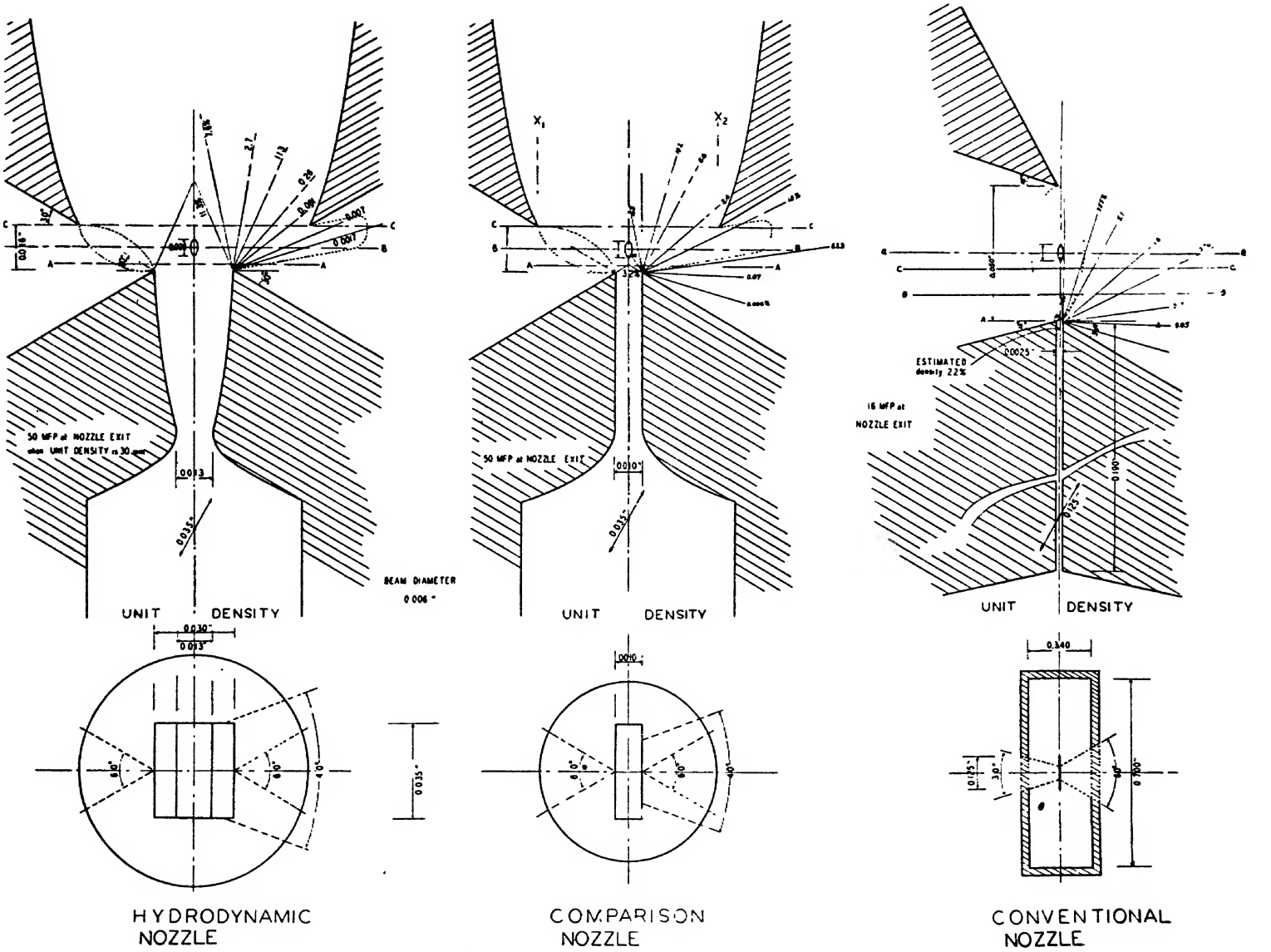


FIG. 2. Vertical and cross sections. Rectangular nozzles designed for two-dimensional flow. The computed isodensity lines just above the exits are given in percentages of the density present in the reservoir. The circle 0.006" diameter shows the position and relative dimension of the electron beam. The electrons proceed from left to right in the figures. The scale for the cross section of 2c is smaller than that of the other figures.

whereas, if liquid air cooling is used, process C_{2b} becomes negligible, and

$$E_{ab}{}^{la}=\frac{C_1+C_3+C_4}{C_1+C_{2a}+C_3+C_4}.$$

The beam absorption data mentioned above are probably more valuable as a means of obtaining an estimate of the extent of multiple scattering relative to the total scattered intensity, once the theoretical relation between these quantities has been established.

From the above discussion it is apparent that an estimate of the merit of a nozzle with respect to sample distribution alone or to multiple scattering alone can be obtained for any specified set of operating conditions as functions of time.

Before these measurements are discussed in detail, it is well to point out that a third definition of efficiency as a function of s, can be obtained from photographic data. This is

$$E_s=\frac{(C_1)_r+(C_3)_r+(C_4)_r}{(C_1)_r+(C_2)_r+(C_3)_r+(C_4)_r},$$

where $(C_n)_r$ is the amount of C_n, integrated over the time of exposure, appearing per unit area on the photographic plate at that radial distance r which corresponds to s for an ideal sample.

DESCRIPTION OF NOZZLES

The manner in which gas flows into the electron diffraction camera depends on the design of the nozzle. In these experiments the three nozzles illustrated in Fig. 2 were tested. First is the so-called "hydrodynamic" nozzle designed by Dr. A. Kantrowitz, of the Graduate School of Aeronautical Engineering of Cornell, to have supersonic flow; the second is referred to as the "comparison" nozzle and has throat dimensions similar to those of the hydrodynamic nozzle but a parallel walled flow channel; the third is the "conventional" nozzle which also has a parallel walled flow channel of about the same cross-sectional area as that of the comparison nozzle but with ratio of width perpendicular to the electron beam to length parallel to the electron beam equal to about 50.

In Fig. 3 is shown the computed gas density destribution across the mouth of the nozzles at three distances from the outlet, *A–A*, *B–B*, and *C–C*. For the conventional nozzle a further line was computed along *Q–Q*. These graphs show the degree of spreading of the gas from the nozzle mouth, and it is clear that the hydrodynamic nozzle tends to give the best directed flow of gas while the conventional nozzle acts more as a source of gas effusion. In some experiments to be described later a *constriction* was placed in the entrance to the *comparison* nozzle, so that the rate of gas flow was reduced and was more along the order of the conventional nozzle.

It is reasonable to expect that little gas should escape from the side ports (through which the electron beam passes) of the hydrodynamic nozzle, but more from the comparison nozzle, and a relatively large amount from the conventional nozzle. This is really a question of the components of molecular velocity perpendicular to the direction of the electron flow. By measuring the flow rates for the three nozzles under similar conditions, and comparing the observed rates with those calculated from flow formulas, we obtained a rough estimate of the pressure drop along the nozzle. Comparison of the estimated pressure drop with that obtained from the inlet and outlet pressures used in the experiments, gave a measure of the number of molecular collisions on the side walls, as the gas proceeds along the nozzle and thus an estimate of the components of velocity perpendicular to the nozzle axis.

It was shown that the hydrodynamic nozzle had virtually no pressure drop along its length; and only about 5 percent of the applied pressure differential

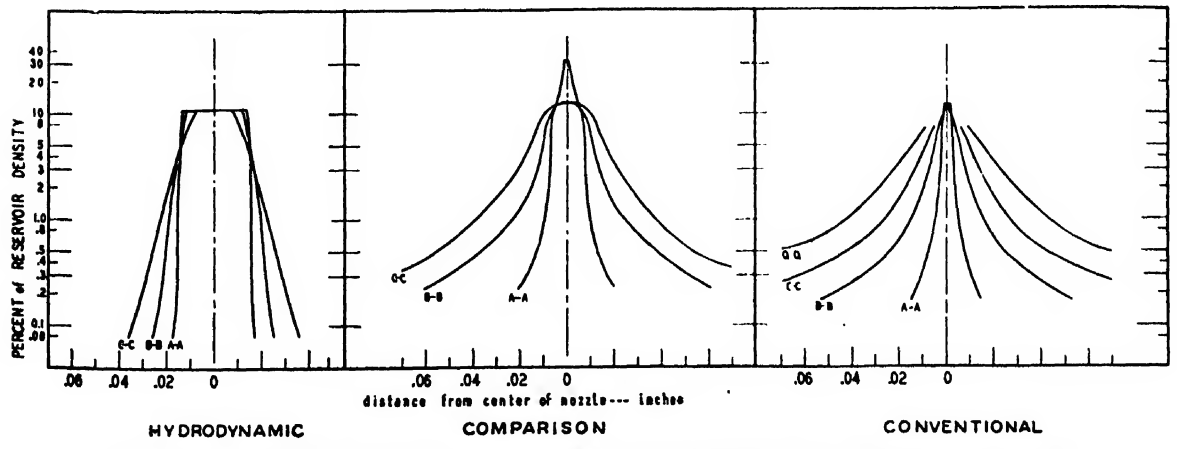


FIG. 3. Computed densities along the ray lines *A–A*, etc. The values for Fig. 2a are reliable; those for 2b are approximate; those for 2c are little better than guesses.

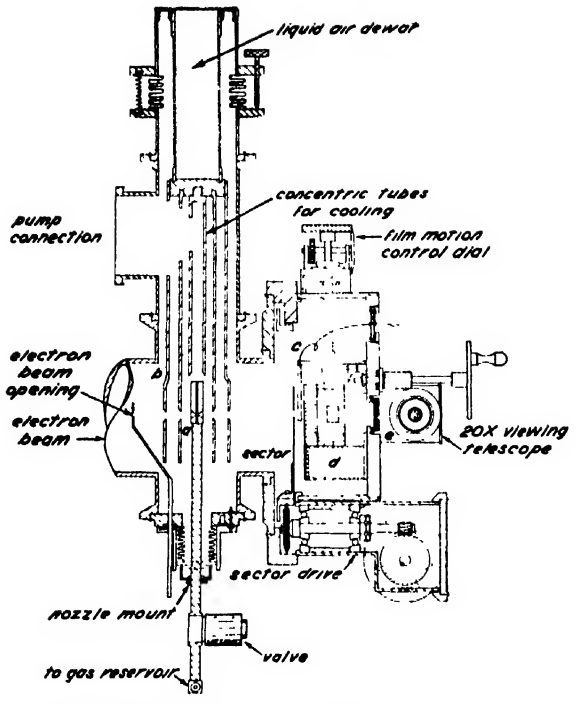


FIG. 4. Side view of electron diffraction apparatus, showing nozzle mount (a), concentric tubes for cooling and condensing the sample (b), pump connection, roll film camera, and sector drive. (c) is a Faraday cage connected to a d.c. amplifier for electron beam control; (d) is a tube carrying a fluorescent screen which is pushed down as the film frame is set in place—the screen remains coplanar to the film; (e) is a short range telescope for viewing the focused spot when the fluorescent screen is in the central position.

was taken up along the walls in the case of the comparison nozzle. Hence the gas exits from the nozzle mouth at a density close to that present at the inlet, and most of the pressure differential is dispersed in the region beyond the nozzle. For the conventional nozzle, one-third to one-half of the applied pressure differential was taken up along the nozzle channel and so the gas must leave the nozzle with comparatively large velocity components perpendicular to the direction of flow. The question may be raised as to what fraction of the gas stream intercepts the electron beam as it fans out from the mouth of the nozzle. It can be shown that under usual operating conditions, most of the gas passes through the beam although with the conventional nozzle the molecules on the average remain in the beam longer so that if comparable gas densities in the diffraction region and exposure times were used for the three nozzles, the observed photographic density would be highest with the conventional nozzle but so would the relative extent of multiple scattering and sample diffusion.

Figure 4 shows a cross-sectional view of the nozzle as placed in the camera. Three concentric chimneys which surround the nozzle can be cooled by filling the metal dewar with liquid air. They also serve to lead the injected gas directly away from the nozzle to the diffusion pump.

THE CALCULATION OF E_p AND THE EFFECTS OF SAMPLE DIFFUSION ON RESOLUTION

The equation for the rate of pressure decrement due to flow of gas at pressure P_s from a volume V_s through a nozzle, into a volume V_c at pressure P_c, may be written

$$-\frac{dP_s}{dt}=\frac{kAu}{V_s}(P_s-P_c)=j(P_s-P_c),$$

where A = area of opening in nozzle, u = velocity of sound, k = flow constant for nozzle, $j=kAu/V_s$. If the volume V_c represents an electron diffraction camera being pumped out at a rate proportional to the pressure P_c, we also must have

$$\frac{dP_c}{dt}=\frac{V_s}{V_c}j(P_s-P_c)-lP_c,$$

where l is a constant depending on the speed of the evacuating pumps. Solution of the above system of differential equations leads to the result

$$P_c=2\frac{mP_0}{b}\left[\exp\left(-\frac{m+l+j+b}{2}t\right)-\exp\left(-\frac{m+l+j-b}{2}t\right)\right],$$

where $m=jV_s/V_c$, $b=((m+l+j)^2-4lj)^{\frac{1}{2}}$, and P_0 is the value of P_s at $t=0$. For the case where V_c/V_s and j/l are much larger than one, this expression may be approximated by

$$P_c=(P_0V_s/V_c)(1-e^{-jt})e^{-lt}. \tag{2}$$

This result is the one which is used in the following considerations.

Since the factor $(1-e^{-jt})$ in the above expression represents the fraction of the gas originally in V_s that has passed through the nozzle in time t, the value of Q_{nozzle} in the expression for E_p may be written

$$Q_{\text{nozzle}}=(1-e^{-jt})(2r/W_n)(\pi r/2\mu),$$

where $(\pi r/2\mu)$ is the average time spent by molecules in traversing an electron beam of radius r, when their average component of velocity perpendicular to the beam is μ, while W_n is the width of the nozzle opening.

$$Q_{\text{camera}}=(\pi r^2R/V_c)Z,$$

where R is the electron beam path length through the camera and

$$Z=\int_0^t(1-e^{-jt})e^{-lt}dt.$$

Hence

$$E_p = \frac{(1-e^{-jt})(2r/W_n)(\pi r/2\mu)}{(1-e^{-jt})(2r/W_n)(\pi r/2\mu)+(\pi r^2 R/V_c)Z}. \quad (3)$$

Aside from apparatus constants our equations require the evaluation of the exponential constants "j" and "l." j is a function of the flow rate through the nozzle and l is a function of the rate of evacuation of the camera; hence two sets of measurements were required. To determine j a large fixed volume of gas was allowed to expand through the nozzle into a vacuum, and the pressure of the volume was recorded as a function of time. From the formula $ln(P_2/P_1) = -j(t_2-t_1)$ we get $j=[-ln(\frac{1}{2})]/t_{\frac{1}{2}}$, wherein $t_{\frac{1}{2}}$ is the time at which $P_2=\frac{1}{2}P_1$.

The determination of l was a little more complex, and to do this an ionization gauge was connected to the electron diffraction camera. The output of the ion gauge was fed to a d.c. amplifier and a permanent record obtained with a Brush recording oscillograph. The variation of pressure with time which was thus obtained is shown in Fig. 5 for a typical case. The sharp rise is characteristic of the factor $(1-e^{-jt})$ and the exponential decrease is due to the (e^{-lt}) factor. The value of l can be determined from the curve since the camera pressure at any time is given by $P=(P_0V_s/V_c) \times(1-e^{-jt})(e^{-lt})$. The calculation of all the constants in the efficiency equation is now straightforward and efficiency can be calculated as a function of time.

Pressure measurements were made under two sets of conditions. By means of the cooling chimneys (Fig. 4) we can condense out most gases before they can diffuse throughout the camera. The curves in Fig. 5a are for the case of the chimneys at room temperature, while those in Fig. 5b are for the chimneys cooled with liquid air. The pressure scale is compressed by a factor of about 5, and the time scale by a factor of about 10 in the latter. Efficiencies were calculated as a function of time, with and without the use of liquid air cooling, and the results plotted in Fig. 6. The following may be pointed out:

(a) E_p begins at 100 percent for $t=0$, and drops for increasing t.

(b) The E_p of a nozzle without adequate liquid air cooling is less than that for a nozzle with adequately cooled condensing surfaces around it.

(c) The comparison nozzle shows the greatest E_p of the three for no liquid air cooling—probably due to the fact that the molecules remain in the beam a longer time than in the case of the supersonic nozzle.

(d) The conventional nozzle is least efficient up to 0.6 second regardless of conditions.

(e) It would follow from the efficiency equations that a high pumping rate would favor higher efficiencies. Liquid air cooled surfaces fulfill this requirement to a substantial extent for most gases. Similarly a *small* camera volume would be desirable, or else one so large that the injection of a gas sample would not raise the pressure appreciably in the camera.

The appearance of electron diffraction photographs was correlated theoretically to the value of the efficiency E_p by comparison of the integrated background intensity (atomic contribution) and of the molecular contribution for an extended sample with the corresponding quantities expected from an infinitesimally thin sample. It was found that when no beam trap is inserted close to the principal region of diffraction, and the sector is placed at the plane of the detector, little departure from unit background is introduced (<2 percent) if the sample is relatively concentrated, as in (a), (c), and (d), Fig. 7. However, for a uniformly distributed sample (Fig. 7, curve *e*), the density at small s is 15–20 percent higher than at large s. The computed effect on the molecular terms due to diffusion of the sample is quite marked. Inspection of the curves in Fig. 8a, b, c will show that due to the overlapping of patterns arising at different sample-detector distances (process 2, Fig. 1) the high frequency terms are more readily damped at large "s" than are the lower frequency contributions. As a consequence the unwary investigator will be led to assign too large temperature factors to the long inter-atomic distances. The above integrations were carried out graphically.

THE CALCULATION OF E_{ab} AND THE EFFECTS OF SECONDARY DIFFRACTION ON RESOLUTION

An estimate of E_{ab} was obtained from measurements of the loss of current in the main electron beam after

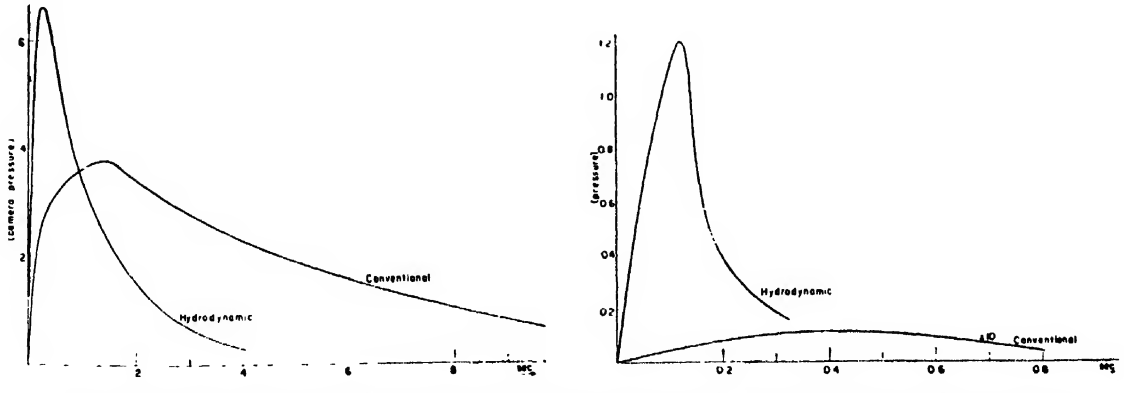


FIG. 5. Pressure transients in the camera. (a) Chimneys at room temperature; (b) Liquid air-cooled chimneys.

it had passed through an injected gas sample under the following set of arrangements:

(a) The main beam was allowed to pass through the nozzle, with the surrounding chimneys at room temperature. Absorption of the beam is by processes 1, 2, 3, and 4.

(b) The nozzle was tilted so that the main beam passed beside it; the nozzle surroundings remained at room temperature. Absorption is by process 2.

(c) The main beam was allowed to pass through the nozzle but the surrounding chimneys were liquid air-cooled. This measures absorption by processes 1, 3, 4, and that part of 2 between nozzle and condensing surfaces (process 2_a).

(d) The main beam was allowed to bypass the nozzle as in (b) but the surrounding surfaces were cooled. Absorption of the beam is by diffraction between the nozzle and the condensing surfaces (process 2_a).

On the basis of measurements such as these one can compute E_{ab} as the ratio of diffraction over the nozzle to the total diffraction at any time during which the sample is in the camera.

The absorption measurements were made with the aid of a d.c. amplifier and Brush recording oscillograph connected to a collecting electrode (cup) of five millimeter diameter placed in the path of the main electron beam at a distance of 19 cm from the nozzle. In Fig. 9b a typical curve of the variation of the main beam current with time is shown for the cases cited above. In curve 1 we have the case of maximum absorption followed by a fairly rapid decay. Case (2) has a much smaller maximum, but it is shifted appreciably to the right indicating, possibly, that gas diffusion around the condensing surfaces and through the camera is a comparatively slow process. Case (3) shows a very sharp rise and rapid decay, for the diffraction taking place over the nozzle. Case (4) failed to give any measurable change in beam current which indicates that absorption of the beam by ambient gas within the cooled condensing surfaces was so small that it was not detected by our instrument. Figure 9a shows a curve of efficiency calculated from the absorption curves.

A fairly good comparison of the nozzles is obtained by quoting $t_{\frac{1}{2}}$ values, i.e., the time for the maximum of the curve to be reduced by 50 percent. See Table I.

In Table II are given the percentages absorption of the beam for the case of three nozzles under various experimental conditions. The values are those of maximum absorption, and the times for maximum values to be reached are also quoted.

The comparison and hydrodynamic nozzles are closely similar and both are more effective on the basis of E_{ab} values than a nozzle of slow gas admittance. If one computes the ratio [diffraction over the nozzle/diffraction elsewhere in the camera] one finds a value of 12 to 13 for the comparison and hydrodynamic nozzles and only about 1.5 for slow gas admittance. With liquid air cooling, the absorption in the camera was not measurable on our apparatus and the E_{ab} values were close to unity in all cases.

It must be pointed out that slower gas flow rates mean lesser absorption of the main beam and as a result less multiple scattering.

In order to use the absorption measurements as a means of estimating the extent of multiple scattering, the following relation, derived in Appendix C, may be used:

$$J_2/J_1=1+[(J_0/J_1)-1]\ln[1-(J_1/J_0)], \qquad (4)$$

where J_0 is the total intensity of the incident main beam, J_1 is the total intensity of scattered radiation, and J_2 is the total intensity of electrons which have been scattered more than once. For values of J_1/J_0 up to about 0.2, J_2/J_1 is fairly well approximated by $0.5\ J_1/J_0$; for J_1/J_0 equal to 0.3 and 0.5, J_2/J_1 equals $0.56\ J_1/J_0$ and $0.61\ J_1/J_0$, respectively. Since J_1/J_0 is measured directly in the absorption experiments, we

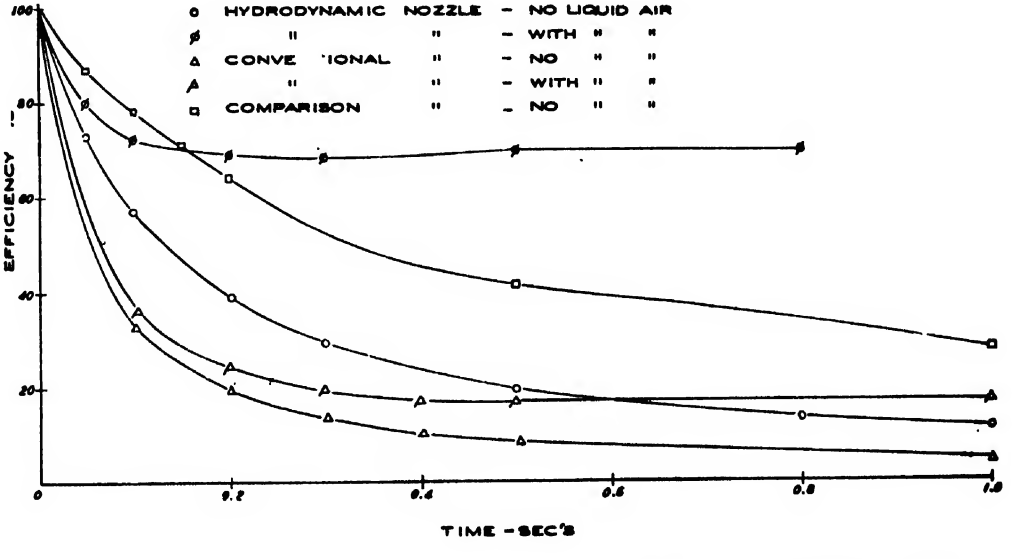


FIG. 6. Nozzle efficiencies (E_p) as a function of the time.

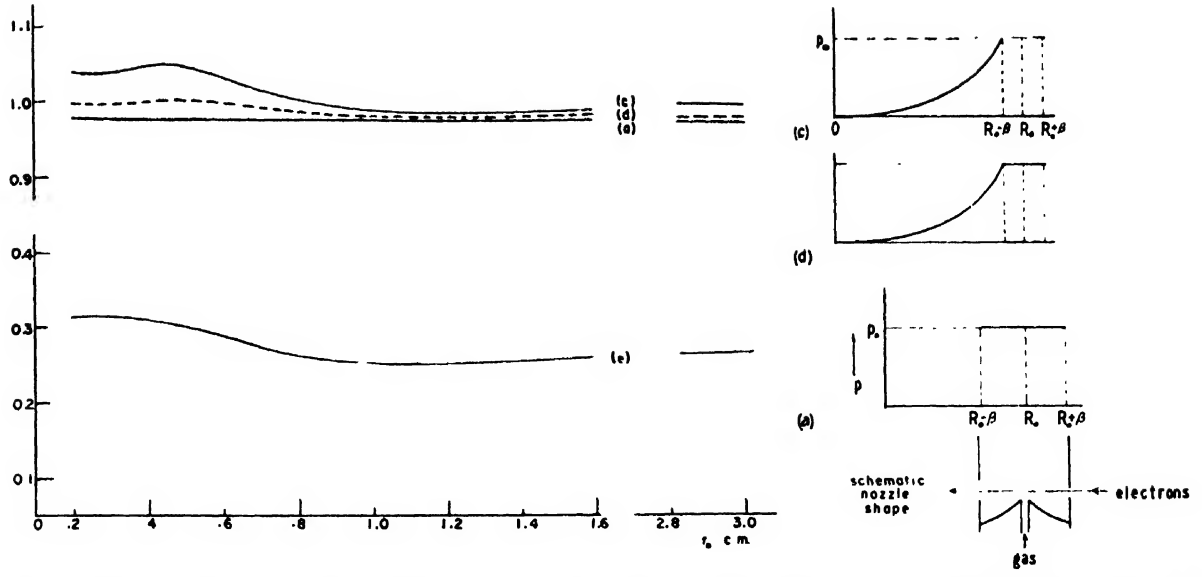


FIG. 7. Relative background intensity curves, for an extended sample compared to that for a sample of the same size concentrated at R_0. The sector is at the plane of the detector, and no beam trap is interposed. The curves correspond to the pressure distributions indicated below, with $R_0=19.33$ cm. $\beta=0.475$ cm. $p=[(p_0\beta^2)/(R_0-R)^2]$ for $R\leqslant(R_0-\beta)$. r_0 is distance along the plate. Curve (*e*) is for a uniform pressure distribution throughout the camera.

have a very simple way of estimating the relative extent of multiple scattering.

For rapid gas admittance (hydrodynamic nozzle), as much as 30 percent of the incident main beam may be absorbed. This means that J_1/J_0 equals 0.30 and hence that the ratio of the amount of multiple scattering to total scattering equals 0.56 times 0.30, or 0.17. For the case of slow gas admittance where the maximum beam absorption may be only 2 percent; we find then that J_2/J_1 equals 0.50 times J_1/J_0, i.e., 0.01, so that the relative extent of multiple scattering has been reduced by a factor of 17 as compared with the case of rapid gas admittance.

In order to correlate these results with the observed appearance of electron diffraction photographs, we must estimate the ratio of the loss of intensity, (I^*-I), resulting from multiple scattering in a ray reaching a particular point on the plane of the detector, to the gain, I_n, of background intensity resulting from the contribution at this point of intensity rescattered from all other rays. In formulating the integrals for this additional background it soon became clear that the general case is very involved; hence we have simplified the discussion by not considering the dependence of the result on the degree of sample extension in the direction of the electron beam. Thus it is assumed that the sample is restricted to a region $(R_0+\beta)$ to $(R_0-\beta)$ directly over the nozzle, and that $\beta/R_0\ll1$ so that rays diffracted at an angle θ originating in any part of the sample all arrive on the plane of the detector at essentially one distance from the central beam. It was also assumed that the gas density is sufficiently low that all of the multiple scattering may be considered to be secondary scattering and that the intensity of the main beam may be considered constant while passing through the sample. Finally, it was found convenient to approximate the variable "s" by $2\pi r/(\lambda R_0)$ where $r/R_0=\tan\theta$. The particular molecule used as an example was carbon tetrachloride, and for it the following empirical relation was fitted:

$$\text{(Atomic contribution + molecular contribution)}\times(2\pi/\lambda R_0)^4 1/s^4=\frac{4280}{(s^2+22)^2}+\frac{310}{(s^2+0.6)^2}. \quad (5)$$

The values of $I_n/(I^*-I)$ which result from these calculations are plotted in Fig. 10. It is apparent that the background intensity is decreased at small s and increased at large s as a result of secondary scattering. It may be demonstrated that the *fractional loss* is independent of s (Appendix A).

To consider a definite example, suppose that the measured absorption of the beam is 20 percent, so that the ratio of the amount of multiple scattering to total scattering is 0.1. Then, assuming for simplicity that at some value of s, say $s=8$, the background intensity I^* that would be obtained if there were no multiple scattering is equal to one, the actual contribution, I, from electrons scattered only once will be $(1.0-0.1)=0.9$, while the contribution I_n resulting from multiple scattering of all other rays will be roughly $2(I^*-I)=0.2$. Hence the total background intensity will be $0.9+0.2=1.1$. Therefore if the ratio [(molecular contribution)/(atomic contribution)] is equal to x under conditions where no multiple scattering occurred, it

will only be equal to $(0.9/1.1)x$ under the actual conditions stated. Indeed, the value will be even less than this because we have not taken into account the fact that rescattering of the molecular contribution will result in a preferential addition of intensity to the valleys of the molecular contribution curve.

Because the background contribution to the total intensity is large compared to the molecular contribution for large values of s even in the ideal case, it is obvious that every effort should be made to reduce the extension, it is apparent that whether the observed background happens to be too high, flat, or too low for small s depends on the balancing of two opposing factors; extension of the sample tends to raise the background (Fig. 7, e) while secondary scattering tends to lower it (Fig. 10). It is quite clear why some observers who failed to consider these complications in nozzle design were misled by the extraneous backgrounds into believing that the atom form factors are necessarily incorrect.

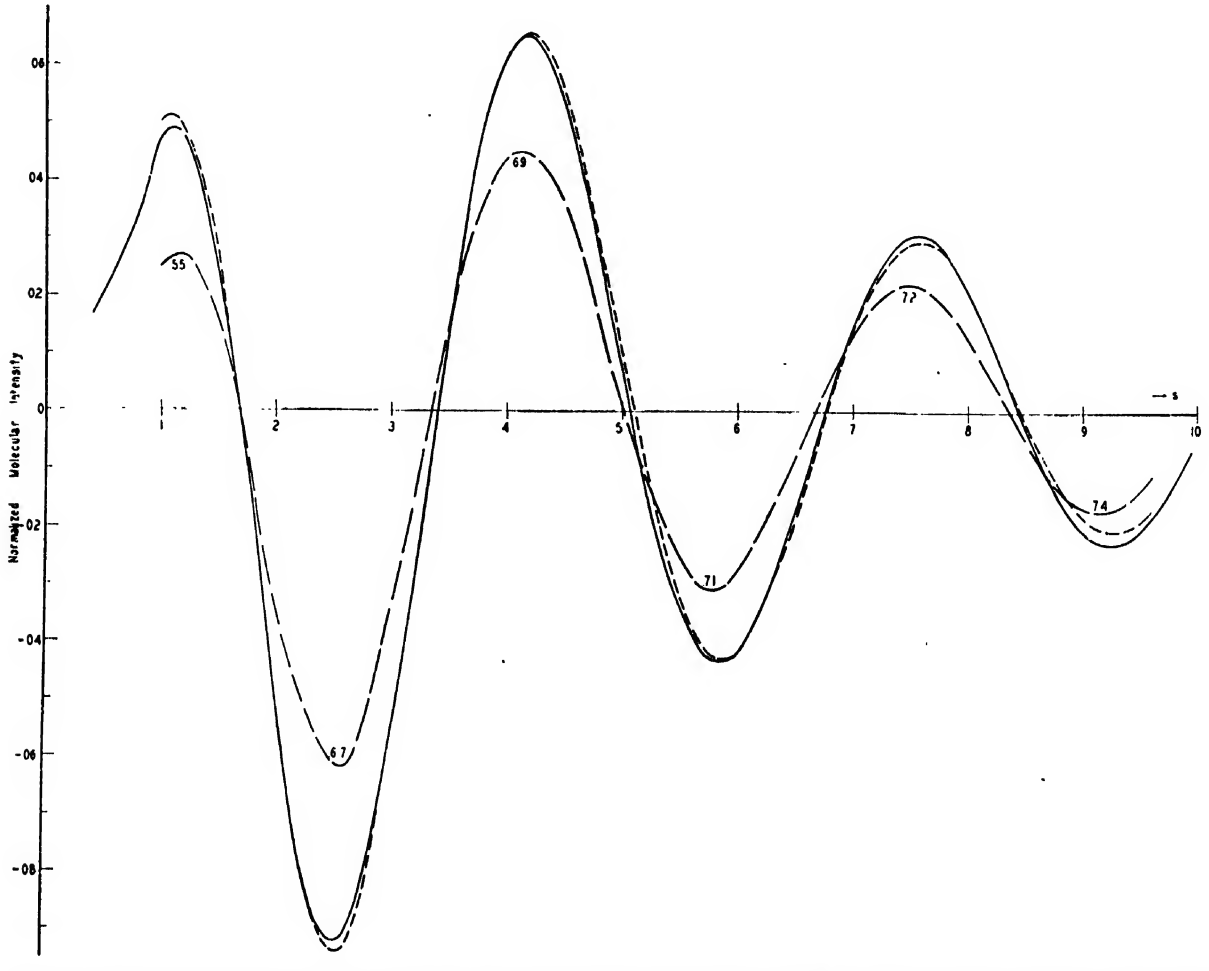


FIG. 8a. Molecular intensity curves for an extended sample using the B–B term of B_2H_6, utilizing a sector and beam trap. Curve Inf. is for an infinitesimally thin sample at $R_0=19.33$ cm; $\lambda=0.06359A°$. $2f_{B^2}e^{-.00415s^2}\times(\sin 1.86s)/1.86s$. In curve a we assumed that the pressure distribution was uniform, extending from $(R_0+\beta)$ to $(R_0-\beta)$, and zero everywhere else. $\beta=0.475$ cm. In curve b the pressure distribution was assumed to be uniform, and to extend from R_0 to $(R_0-\beta)$; i.e., an unsymmetrical nozzle. Note the effect of the beam stop on the positions of the maxima and minima, and on their relative intensities. The numbers on this curve indicate the fractional intensities of the corresponding peaks compared to curve Inf. [The broken line curve should be designated a above the abscissa and b, below. The solid line curve should be marked Inf.]

beam absorption in the nozzle to a low value. To do this requires of course that the sample density be decreased and that the exposure times be correspondingly lengthened to an extent consistent with the limitations imposed by stability of the electron gun, high voltage, pumping rates, etc.

Also, from these results and the ones for finite sample

PHOTOGRAPHIC EVIDENCE OF LOSS OF RESOLUTION

A series of photographic records (i.e., actual electron diffraction pictures) was made using cooled condensing surfaces in some exposures and none in others, and tilting the nozzle so that diffraction by gas in the camera could be measured. We can thereby define an

efficiency, E_s. If the photographic plate is read with a microphotometer and a record obtained, a line can be drawn through the fluctuations of the curve in such a way as to represent the background. Comparison of this background when the beam passes through the nozzle (case of normal diffraction procedure) to the background when the nozzle is tilted and the beam passes through the gas in the camera only, gives

$$E_s=\frac{D_{\text{nozzle}}}{D_{\text{nozzle}}+D_{\text{camera}}}=\frac{(C_1)_r+(C_3)_r+(C_4)_r}{(C_1)_r+(C_2)_r+(C_3)_r+(C_4)_r}.$$

E_s values can be determined for the case of liquid air cooled surfaces around the nozzle and for the surfaces at room temperature. The measurements were made on Eastman Commercial plates when no sector was used, and on Eastman Microfile when a sector was used. Sensitometric curves were determined on both of these emulsions so that the photographic densities could be converted to electron intensities. The photographic densities were measured with a Leeds and Northrup recording microphotometer. CCl_4 vapor was used as a condensable gas throughout the experiments.

E_s values are plotted in Fig. 11 for the hydrodynamic nozzle and for the comparison nozzle with a constriction so that the gas entered slowly. The hydrodynamic nozzle with the surrounding chimneys at room temperature allows a considerable amount of gas to diffuse through the camera so that the efficiency is considerably less than 100 percent. The observed dropping off of the efficiency with decreasing s should be expected because of the fact that the background contributed by that portion of the sample which has diffused through the camera falls off more rapidly with s than that contributed by an ideal sample, and because of the fact that multiple scattering in the dense gas directly over the nozzle lowers the background contributed by it at small s values. Therefore from the standpoint of this latter factor, the apparent increase in efficiency with s is actually an indication that resolution is *reduced* at large s values. With liquid air cooling the efficiency remains at about 99 percent irrespective of the angle of diffraction; this indicates that the liquid air-cooled shields effectively remove the diffracted gas but implies nothing about the seriousness of multiple scattering. It is interesting to note that the nozzle of slow gas admittance (comparison nozzle plus constriction) gives better E_s values than the hydrodynamic nozzle. This may be a result of more efficient operation of the evacuation pumps when the pressure in the camera is kept low, as will be the case if the gas is injected into the camera slowly. With the chimneys liquid air-cooled, efficiencies of about 99 percent were again obtained for slow gas admittance.

The actual resolution for the diffraction patterns can be determined by reading the height of the peaks above the valleys in ratio to the background and comparing this to the theoretical ratio of peak height to background. These figures give us, then, the loss on intensity of the peaks and gain of intensity of the background. That secondary scattering and diffraction by an extended sample do lead to a loss in resolution can be demonstrated in the following simple way. Suppose

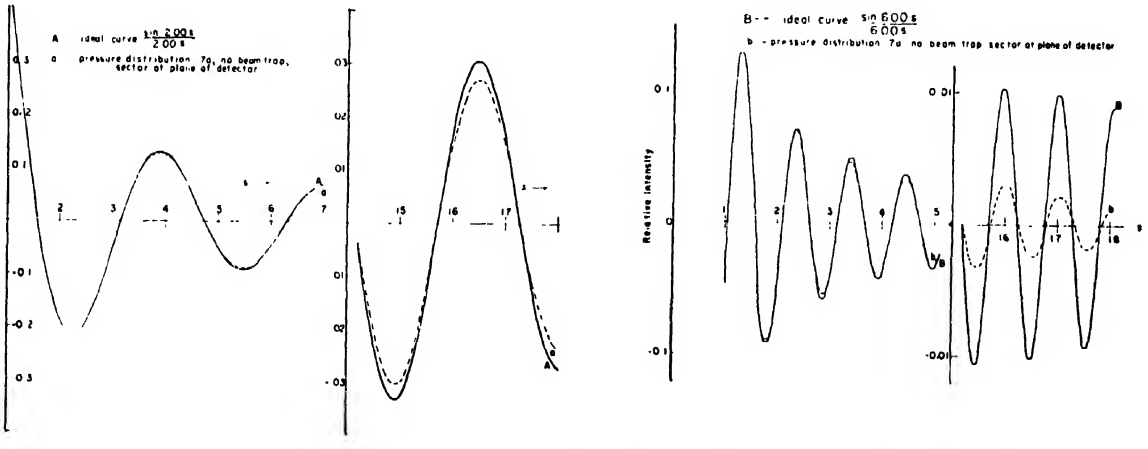
$$I_{\text{theoretical}}(s)=A_1(s)+A_2\sin\gamma s$$

with $A_2<A_1(s)$, which is a slowly varying function. Let the fraction $(1-\beta)$ be lost through multiple scattering, and let the total gain in background be $[B_1(s)+B_2(s)]$. Then

$$I_{\text{observed}}(s)=[B_1(s)+B_2(s)]+\beta[A_1(s)+A_2\sin\gamma s].$$

A measure of the resolution is the magnitude of the observed "relative fluctuation," i.e.,

$$\frac{2\beta A_2}{[B_1(s)+B_2(s)]+\beta A_1(s)}\left[\frac{2A_2}{A_1(s)}\right]^{-1}\cong 1-\frac{B_1(s)+B_2(s)}{\beta A_1(s)}.$$



FIGS. 8b and 8c. No beam trap correction was inserted in these integrations, and the sector was placed in the plane of the detector. The pressure distributions [they should be marked 8a rather than 7a in both parts of the figure] postulated were the same as for Fig. 8a, curve *a*.

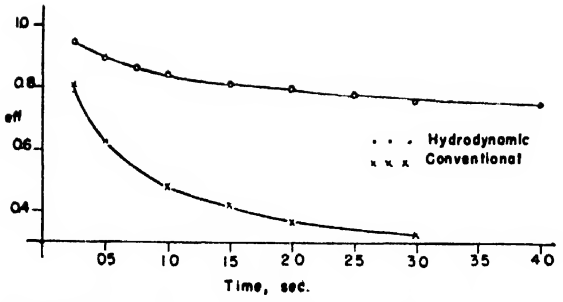


FIG. 9a. Nozzle efficiencies (E_{ab}) as a function of the time.

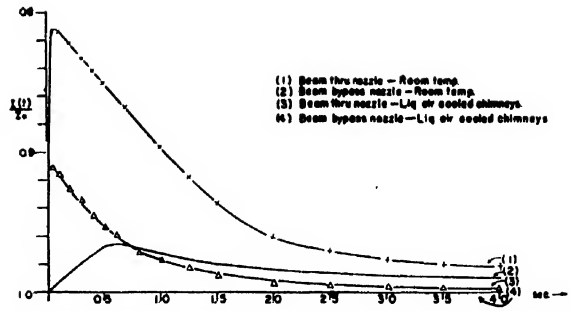


FIG. 9b. Records of beam current absorption curves.

Thus, the resolution is diminished by a factor which is inversely proportional to the fraction of each ray which gets by without suffering secondary scattering, and directly proportional to the gain in background. Further, it is probable that $B_2(s)$, which arises from the secondary scattering of the A_2 term, is a harmonic function of s, but with a phase which does not match that of $\sin\gamma s$; this effect will introduce a further decrease in resolution.

By measuring the beam absorption and computing the ratio of peak height to total intensity (at the 7th peak for CCl_4) for a series of inlet pressures we obtained the results listed in Table III. This table shows very clearly the effect of the rate of gas admittance on the resolution since we reduced our efficiency by a factor of 3 for increase of gas pressure in the reservoir by a factor of 7 or so. As the beam absorption increases from 8 percent to 16.7 percent, the resolution, as measured, drops by a factor of almost 2. Actually, our beam absorption figures are smaller than they should be, for the Faraday cage introduced to remove the main beam between the nozzle and the detector must be larger than the beam in order to avoid edge scattering. Hence a part of the diffracted pattern is trapped with the main beams when gas is admitted.

To conclude this section it would be in order to say a word about the effects on resolution of imperfect focus at the plane of the detector. A large scale microphotometer record was prepared from an etched gold foil diffraction pattern. From this the distribution of intensity over a ring, $K(s_0-s)$, was determined, due not only to the finite size of the beam but also to the finite size of the slit in the L and N microphotometer. The observed net effective width at half maximum for the (111) ring was 0.03 cm,[4] with $R_0=19.33$ cm. By comparing the integrated $\int K(s_0-s)I(s)ds$, with the ideal $I(s)$ gas diffraction patterns, it was found that:

(1) there is no lateral shifting of the maxima and minima;
(2) the ratio of the observed peak heights to the theoretical peak heights is less than unity, i.e., there is a loss of resolution;
(3) This ratio is independent of s, but decreases with increasing interatomic distances. This effect, similar to that arising from extension of the sample, also leads to the assignment of false temperature factors.

Interatomic distances	1.86	5.58	7.44
Ratio $\frac{\text{integrated}}{\text{ideal}}$	0.97	0.84	0.72.

Actually, a slight dependence of resolution on s is introduced in most electron diffraction apparatus through the use of a flat plate instead of one curved to the focal circle. However, the additional broadening is quite small (at $s=23$ the effective increase in half-width of the diffraction peak is less than 1 percent).

TABLE I. Values of $t_{\frac{1}{2}}$ for beam absorption measurements.

Nozzle	No liquid air		Liquid air	
	In beam	Out of beam	In beam	Out of beam
Hydrodynamic	0.5 sec.	1.3 sec.	0.36 sec.	0.00 sec.
Comparison	0.57 sec.	2 sec.	0.45 sec.	0.00 sec.
Comparison +constriction	6 sec.	6 sec.	2 sec.	0.00 sec.

A MOLECULAR BEAM NOZZLE

From the data presented above, one is led to speculate on the desirability of using a molecular beam in order to avoid the effects of secondary scattering and spreading of the gas through the camera. The principal difficulty in using a conventional molecular beam is that the gas densities are not sufficiently high to avoid making unmanageably long exposures. However, it is

TABLE II. Absorption of main electron beam and time to maximum absorption.

Nozzle	No liquid air		Liquid air	
	In beam	Out of beam	In beam	Out of beam
Hydrodynamic	31 % at 0.08 sec.	2.1% at 0.37 sec.	27 % at 0.06 sec.	0.0%
Comparison	35 % at 0.10 sec.	2 % at 0.5 sec.	25 % at 0.10 sec.	0.0%
Comparison +constriction	2.4% at 0.80 sec.	1.1% at 1.5 sec.	1.6% at 0.1 sec.	0.0%

[4] Later, measurement of a trace of the direct beam showed that at half maximum, it is 0.0035 cm wide, so that the effects computed below are much larger than those actually present. The diffraction broadening due to our gold foil is approximately 10 times the diameter of the focused spot.

not necessary that a molecular beam used for electron diffraction work have the degree of collimation required for other types of studies. A possible design for a nozzle satisfactory for electron diffraction is indicated below. Experimentally we found that in order to obtain sufficient photographic density in the diffraction photographs ($D=0.2$ at $s=25$ for microfile film), there must be approximately $(5\times10^{+14})/ti$ molecules in the electron beam per square cm of beam cross-sectional area, where t is the duration of the exposure in seconds, and i is the total beam current in microamperes.

Let us begin by supposing that a source slit S_1 (see Fig. 12) of dimensions 0.01×0.1 cm is used. The small dimension of such a slit determines the maximum gas sample density which may be used without obtaining an appreciable number of collisions in the gas stream beyond S_1. The maximum density is the one at which the gas molecules have a mean free path equal approximately to the small dimension of the slit. Thus for a slit 0.01 cm wide, the maximum sample pressure which one may safely use is about 0.5 mm; or N, the number of molecules per cc in the reservoir, is about 1.5×10^{16}.

The gas density which appears directly above S_2, where the electron beam passes by, is approximately[6] $N'=Na/4\pi l^2$ where N' is the gas density in molecules per cc at S_2, a is the area of the source slit S_1, and l is the separation of S_1 and S_2. However, what is actually required is not N', but N'', the number of molecules in the electron beam per square cm of beam cross section; this quantity may be obtained simply by multiplying N' by that dimension of S_2 which lies along the direction of the electron beam. Therefore if this dimension is 0.1 cm, we find that

$$N''=\frac{1.5\times10^{16}\times1.0\times10^{-3}}{12.5l^2}\times0.1=\frac{1\times10^{11}}{l^2}.$$

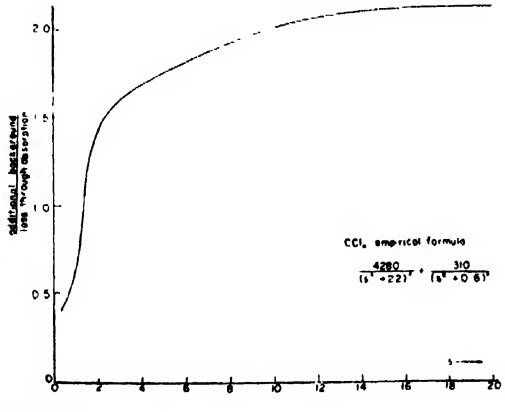


FIG. 10. Computed values for the ratio

$$\frac{\text{Addition background due to secondary scattering}}{\text{Loss in background due to absorption}}$$

as a function of s.

[6] R. G. J. Fraser, *Molecular Rays* (Mcamillan Company, Inc. New York, 1931), p. 14.

TABLE III. Observed resolution.

CCl_4 press	Absorption % (as measured using 5 mm collector)	Peak height / Total intensity ×100	Resolution (%)
—	0	12.0 (theoretical)*	100
16.5 (−12°C)	8	11.5	96
32.9 (0°C)	16.7	6.8	57
119.5 (26°C)	46.5	3.9	33

* Note: the theoretical curve was constructed using the temperature factors $e^{-0.0018s^2}$ and $e^{-0.0021s^2}$ for C−Cl distances and Cl−Cl distances, respectively.

From this relation the approximate exposure times required to give satisfactory electron diffraction photographs can be calculated as a function of l. Assuming that the main beam current i equals one microampere, we thus find that the time of exposure t equals 2×10^4, 800, and 110 seconds for l equal to 2, 0.4, and 0.15 cm, respectively.

From the standpoint of short exposure times one would like l to be small. On the other hand, if l is made very small, not only will the angular divergence of the beam be large, but also molecules which hit the bottom edges of S_2 will require too many additional bounces from the surfaces of S_1 and S_2 to reach the wide throat of the pump; the greater the average number of bounces required, the greater will be the probability for any one of these molecules to go through S_2 along an undesirable path rather than to the diffusion pump. For this reason we wish to select a compromise value—to make S_2 wide enough to allow a good fraction of the molecules leaving S_1 to go directly through S_2, but still to maintain molecular beam characteristics. The practical limit is determined by the maximum sample extension which is to be allowed. With $l=0.15$ cm and S_2 0.1×0.1 cm, the total angular divergence will be about 42° so that the maximum sample extension along the direction of the electron beam at a distance 0.1 cm above the bottom of S_2 will be 0.2 cm, a not unreasonable value. Roughly 55 percent of the molecules leaving S_1 will go directly through S_2 in this case.

Table IV summarizes some of the characteristics and requirements of this type of molecular beam nozzle, and Fig. 12 illustrates its physical outlines.

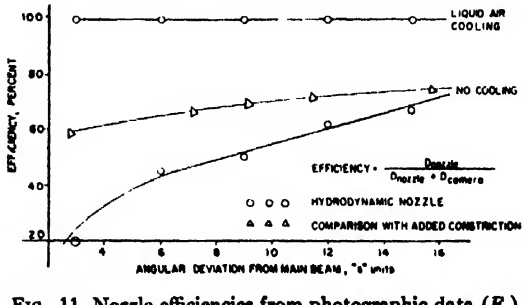


FIG. 11. Nozzle efficiencies from photographic data (E_s), as a function of s.

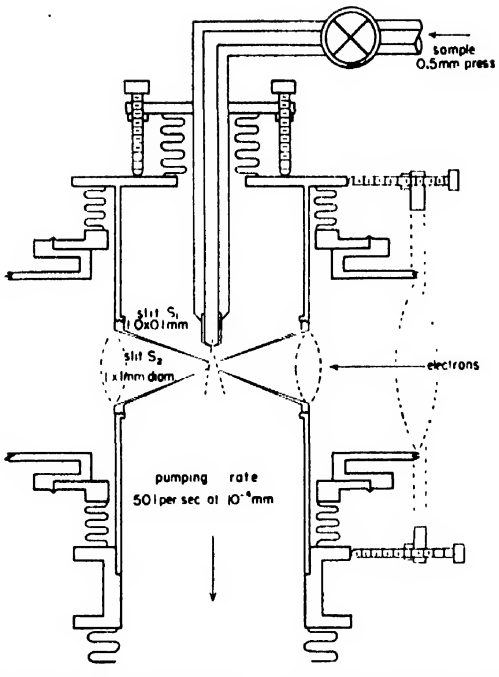


FIG. 12. Proposed molecular beam nozzle and pumping section.

If the gas densities achieved by this nozzle are found to be too low, or the angular divergence too great, a source of high gas density has been designed by Dr. A. Kantrowitz[6] of the Graduate School of Aeronautical Engineering at Cornell, utilizing supersonic flow. Under comparable conditions a molecular beam intensity about 75 times that obtainable with standard molecular beam geometry can be produced. However, due to its added complexity and inefficient sample utilization it would not be advantageous unless the factors of higher gas density, and lower effective temperature of the gas (approximately 72°K for a Mach number 4) were necessary for the investigation in hand.

TABLE IV. Properties of modified molecular beam nozzle.*

Property	Calculated value	
Dimension of slit S_1	0.01×0.1 cm	
Dimension of slit S_2	0.1×0.1 cm	
Pump capacity (on nozzle—Fig. 12)	50 l/sec. at 1×10^{-4} mm	
Molecules in electron beam per unit area of electron beam	6×10^{11}	
Molecular velocity	5×10^{4} cm per sec.	
Effective molecular temperature	300°K	
Consumption of sample	3×10^{-7} moles/sec.	
Exposure time and total beam divergence	time	angle
S_1-S_2 distance 2 cm	5.5 hrs.	3°
4 mm	13 min.	14°
1.5 mm	2 min.	42°

* With beam current increased by a factor of 5 (to 5 microamperes) the exposure time in the last case will be reduced to 24 seconds.

[6] J. Grey and A. Kantrowitz, Phys. Rev. **75**, 1331 (1949). For details refer to Master's Dissertation submitted by Mr. Grey to the Graduate School of Aeronautical Engineering at Cornell University (1949).

DISCUSSION AND CONCLUSIONS

We have attributed the lack of resolution of the diffracted electron image on the detector to failure of assumptions (2), (3), and (4) listed in the introduction. The data obtained may be summarized as follows:

Pressure Measurements

(i) $E_p(l)$

$$= \frac{\text{molecules coming from the nozzle}}{\text{molecules from nozzle}+\text{molecules from camera}}$$

(ii) Better efficiencies are given by the hydrodynamic or comparison nozzles than the conventional, with or without the use of liquid air.

(iii) Efficiencies are always higher for shorter exposures, and for high camera pumping speeds.

(iv) It is desirable to beam the gas through the camera and directly into a diffusion pump.

Beam Absorption Measurements

(i) $E_{ab}(l)=\dfrac{\text{beam absorption over the nozzle}}{\text{beam abs. over nozzle}+\text{beam abs. in camera}}$

This is greater for the hydrodynamic or comparison nozzles than for the nozzle of slow admittance by a factor of 50 percent when no liquid air cooling is used.

(ii) Beam absorption with the hydrodynamic or comparison nozzles may be as great as 30 percent, and secondary scattering within the nozzle will be serious.

(iii) Liquid air-cooled surfaces *do not* reduce secondary scattering over the nozzle exit.

Photographic Measurements

(i) The efficiency is an integrated value over the exposure time, being defined:

$$E_s=\frac{\text{Photographic Density from gas out of nozzle}}{\text{P.D. out of nozzle}+\text{P.D. with beam by-passing nozzle}}.$$

(ii) At room temperature the efficiencies become strongly dependent on the angle of diffraction.

(iii) With liquid air-cooled surfaces around the nozzle the efficiencies are close to 100 percent.

(iv) Secondary scattering may be serious; it is dependent on scattering angle; and the shape of the background can be appreciably changed with density of the sample gas (probably due to the differences in secondary scattering). When absorption of the main beam exceeds about 10 percent, resolution is considerably impaired.

In the analysis of secondary scattering it was shown that an angular dependence existed, although secondary scattering was not dependent on the sample distribution. The ratio of [gain in background/loss of primary diffraction] rises to 1.6 (at about $s=3$), and continues

to increase to a value about 2 for larger values of s. The secondary scattering is chiefly a function of N, the number of molecules available to the beam. The conventional nozzle, because of its slow gas admittance, will give rise to the least multiple scattering but allows a great deal of sample extension to occur. The hydrodynamic nozzle, under optimum conditions, gives a much better defined gas stream but for hydrodynamic principles to apply, a gas density so high that multiple scattering becomes very serious must be used and hence this nozzle has little if any advantage over the comparison nozzle, especially if liquid air-cooled chimneys which reduce the seriousness of sample extension can be used. As a means of avoiding to a large extent the effects of both sample extension and multiple scattering, a design for a modified molecular beam nozzle suitable for electron diffraction work is offered. With any type of nozzle, the escaping gas should be able to flow in a straight path directly into a diffusion pump of high speed.

With proper nozzle and camera design it should be possible to achieve nearly maximum resolution in the electron diffraction patterns of gases.

ACKNOWLEDGMENT

Several grants made the completion of this investigation possible. We wish to thank sincerely: The Canadian Defence Research Board—a fellowship for R. B. Harvey; The Office of Naval Research—a research assistantship for F. A. Keidel (under contract No. N6-ori-213, Task I NR 052-040); The Guggenheim Memorial Foundation—a fellowship for S. H. Bauer. All of us are indebted to Professor A. Kantrowitz for his assistance with nozzle designs.

APPENDIX A. ANGULAR INDEPENDENCE OF ABSORPTION OF DIFFRACTED RAYS

The statement that the absorption of the diffracted rays is independent of s may be justified on the assumption that the actual gas distribution in the vicinity of the nozzle is cylindrically symmetric about the axis along which the main stream of gas flows into the camera, while the incident electron beam enters along a diameter of this cylinder. Then, if J_0 is the intensity reaching a primary beam collector when no sample is present, and $N(R, t)$ is the distribution in the number of molecules per unit volume, as a function of distance from the gas stream axis and time, the primary beam intensity, observed when a sample is present, will be

$$J(t)=J_0\exp\left[-\pi\sigma^2\int_0^{R_0+\xi}N(R,t)dR\right], \qquad \text{(A1)}$$

where σ is the electron-molecular collision radius, and is in a sense defined by the above absorption equation, and $(R_0+\xi)$ is the maximum extension of the sample, so that the limits 0 to $(R_0+\xi)$ include all the gas molecules which intercept the beam, from the last aperture in the magnetic lens to the camera.

If there were no higher order scattering, the intensity per unit area of collector at some point r in the plane of the observer, measured from the undeflected ray, would be

$$I^*(r/R_0,t)=J_0B^2\int_0^{R_0+\xi}N(R,t)\frac{(\text{atomic}+\text{molecular})}{R^2[s(r/R)]^4}dR, \qquad \text{(A2)}$$

where

$$B^2=\left[\frac{8\pi^2me^2}{h^2}\right]^2;\quad s(r/R)=\frac{4\pi}{\lambda}\sin\frac{\theta}{2};\quad \theta=\arctan\frac{r}{R}.$$

Equations (1) and (A2) are identical except that the former is valid only if all the N molecules are located at R_0, whereas in (A2) we have permitted the sample to be distributed as per $N(R, t)$. Thus it gives the intensity for all rays diffracted by an extended sample assuming no secondary scattering. However, higher order scattering introduces an effective absorption so that the actual intensity of electrons scattered only once will be

$$I(r/R_0,t)=B^2\int_0^{R_0+\xi}J_0\exp\left[-\pi\sigma^2\int_R^{R+\xi}N(R,t)dR\right]N(R,t)\times\frac{(\text{atomic}+\text{molecular})}{R^2s^4}\exp\left[-\pi\sigma^2\int_0^{R}N(R,t)dR\right]dR. \qquad \text{(A3)}$$

The first exponential factor allows for the attenuation of the primary ray in proceeding from $(R_0+\xi)$ to R; the second allows for the attenuation of the diffracted ray in proceeding from R to the plane of the detector (at 0). Hence, subject to the assumption that *the angles of diffraction are not large*,

$$I(r/R_0,t)=I^*\exp\left[-\pi\sigma^2\int_0^{R+\xi}N(R,t)dR\right]. \qquad \text{(A4)}$$

Thus *no selective loss* is to be expected for any ray, be it primary or diffracted. Similarly, both the atomic and molecular contributions will be reduced by the same fraction.

APPENDIX B. COMPUTATION OF COLLISION CROSS SECTION

The absorption coefficient $\pi\sigma^2$ may be estimated in the following way: Circumscribe around a sample located in its entirety at R_0, a sphere of radius R_0. Then an element of area on that sphere is given in terms of the solid angle $\Delta\theta\Delta\phi$ by $[R_0\sin\theta\Delta\phi R_0\Delta\theta]$. The intensity of electrons scattered into this element of solid angle by $N_0(t)$ molecules (per unit area of the incident beam),

$$j(\theta,t)=J_0B^2N_0(t)\frac{(\text{atomic}+\text{molecular})}{s^4}\sin\theta\Delta\phi\Delta\theta. \qquad \text{(B1)}$$

Hence, the fraction lost from the primary ray, for a *sample of infinitesimal thickness*,

$$\frac{dj(t)}{J_0}=2\pi\int_0^{\pi}\frac{j(\theta,t)d\theta}{J_0}=2\pi B^2N_0(t)\int_0^{\pi}\frac{(\text{atomic}+\text{molecular})}{s^4}\sin\theta d\theta=\pi\sigma^2N_0(t). \qquad \text{(B2)}$$

Hence we have

$$\pi\sigma^2=2\pi B^2\left(\frac{\lambda}{2\pi}\right)^2\int_0^{4\pi/\lambda}\frac{(\text{atomic}+\text{molecular})_s}{s^2}ds. \qquad \text{(B3)}$$

The above integral may be evaluated through the use of Bewilogua's tables.* For CCl_4 this reduces to 1.71×10^{-17} cm^2, provided one makes appropriate assumptions as to the atom form factors at small angles. We also measured this quantity by making preliminary estimates of the magnitude of

$$\int_0^{R_0+\xi}N(R,t)dR \quad \text{and substituted in (A1).}$$

This leads to 2.1×10^{-17} cm^2 for the value of σ^2 for CCl_4.

We have thus demonstrated that the *fractional loss* due to higher order scattering is independent of the angle, and is directly measurable through the attenuation of the primary beam. For CCl_4 under conditions which are typical for most electron diffraction experiments, we estimated the average secondary and higher order scattering to be, after writing $N(R, t)=N_0(t)N_1(R)$,

$$\frac{1}{t_x}\int_0^{t_x}\frac{(I^*-I)}{I^*}dt=\frac{\pi\sigma^2}{t_x}\int_0^{t_x}N_0(t)dt\int^{R_0+\xi}N_1(R)dR=(4\text{ to }11)\times10^{-2} \qquad \text{(B4)}$$

* L. Bewilogua, Physik. Zeits. 32, 740 (1931). See also W. Heisenberg, Physik. Zeits. 32, 737 (1931).

where t_s is the exposure time. These are values which we believe are typical for most current electron diffraction experiments, and they certainly are greater than the error we wish to allow.

APPENDIX C. THE RATIO OF TOTAL MULTIPLE SCATTERING TO TOTAL SCATTERING

Divide interval 0 to $(R_0+\xi)$ into n slices of thickness $(\Delta R)_m$ so that each slice contains the same number of molecules, c/n, per unit area. Set $\pi\sigma^2=K$.

Then the amount of diffraction in the main beam occurring in slice m between R_m and R_{m+1} is

$$J_1(m)=J_0\left\{\exp\left[-K\int_{R_m+\Delta R_m}^{R_0+\xi}NdR\right]-\exp\left[-K\int_{R_m}^{R_0+\xi}NdR\right]\right\}$$
$$=J_0\left\{\exp\left[-K\frac{c(n-m)}{n}\right]-\exp\left[-K\frac{c(n-m+1)}{n}\right]\right\}.$$

The total amount of diffraction is therefore

$$J_1=\sum_{m=1}^{n}J_1(m)=J_0\{1-\exp[-Kc]\}.$$

From this we find at once that the total amount of intensity rescattered from the rays originating in slice m is given by*

$$J_2(m)=J_0\left\{\exp\left[-K\frac{c(n-m)}{n}\right]-\exp\left[-K\frac{c(n-m+1)}{n}\right]\right\}\cdot\left\{1-\exp\left[-K\int^{R_m}NdR\right]\right\}$$

and the total scattering from all rays originating in all the slices is

$$J_2=\sum_{m=1}^{n}J_2(m)=\sum_{m=1}^{n}J_0\left\{\exp\left[-K\frac{c(n-m)}{n}\right]-\exp\left[-K\frac{c(n-m+1)}{n}\right]\right\}\cdot\left\{1-\exp\left[-K\frac{m}{n}C\right]\right\}$$
$$=J_1+\sum_{m=1}^{n}J_0\left\{-\exp[-Kc]+\exp\left[-Kc\frac{n+1}{n}\right]\right\}$$
$$=J_1+nJ_0\{-\exp[-Kc]+\exp[-Kc(n+1/n)]\}$$
$$=J_1+nJ_0\{-1+\exp[-Kc/n]\}\exp[-Kc].$$

Now the value of

$$\lim_{n\to\infty}n\{e^{-Kc/n}-1\}=\lim_{n\to\infty}n\{-Kc/n\}=-Kc.$$

Hence

$$J_2=J_1+J_0\cdot(-Kc)\exp[-Kc]$$
$$=J_1+(J_0-J_1)(-Kc)$$
$$=J_1+(J_0-J_1)\ln[(J_0-J)/J_0]$$

or

$$J_2/J_1=1+[(J_0/J_1)-1]\cdot\ln[1-(J_1/J_0)].$$

* Ignores slight increase in path length because of angle and any change in gas density as a function of the distance *perpendicular* to the electron beam. These will be very small effects ordinarily since most of the intensity appears at small angles.

Studies with an Aerodynamically Instrumented Shock Tube*†

A. Hertzberg and A. Kantrowitz
Cornell University, Ithaca, New York

(Received April 6, 1950)

A shock tube has been set up at Cornell University to begin the study of two-dimensional wave propagation. Pressure waves initiated by the shock itself were used to trigger the spark source for the schlieren flow-observation system, to time the instant of exposure and to measure the strength (speed) of the shock waves. These techniques were employed to study the propagation of a wave in a converging channel. Fair agreement between the one-dimensional method of characteristics and these experiments was obtained.

INTRODUCTION

A SHOCK tube is a device in which simple normal shocks are produced by the sudden bursting of a diaphragm separating regions of higher and lower pressure gases. The compression waves produced in the lower pressure region rapidly steepen into a shock which propagates with a front normal to the axis of the channel. This device has proven convenient for many studies of shock wave phenomena.[1-5]

A shock tube has been constructed in the Graduate School of Aeronautical Engineering at Cornell and has been used for studies of the limitations of the one-dimensional characteristic theory.[6,7] Most previous shock tubes have used electronic instrumentation for measuring the strength of the shock wave generated and to trigger the spark at a desired time. The central idea in this work has been to simplify the instrumentation by using aerodynamic wave phenomena. Thus, weak waves traveling at nearly sonic speed were photographed simultaneously with the phenomena under study and

* This work has been largely supported by the ONR.
† Permanent equipment used in this research was furnished by the J. Carlton Ward Fund for Aeronautical Research of Cornell University.

[1] W. Payman and W. C. F. Shepherd, "Explosion waves and shock waves—The disturbances produced by bursting diaphragms with compressed air," Proc. Roy. Soc. A186, 293 (1946).

[2] F. B. Harrison and W. Bleakney, "Remeasurement of reflection angles in regular and Mach reflection shocks," Princeton University (1947).

[3] Huber, Fitton, and Delpino, "Experimental investigation of moving pressure disturbances and correlation with one-dimensional unsteady flow theory," N.A.C.A. T.N. 1903 (1949).

[4] C. Donaldson and R. D. Sullivan, "The effect of wall friction on the strength of shock waves in tubes and hydraulic jumps in channels," N.A.C.A. T.N. 1942 (September, 1949).

[5] G. R. Cowan and D. F. Hornig, "Technical report No. 1—The thickness of a shock front in air," Metcalf Research Laboratory, Brown University (April 6, 1949).

[6] A. Kantrowitz, "Heat engines based on wave processes" (Grad. Sch. of Aero. Eng.), Cornell University (1948). Presented at the annual meeting of the A.S.M.E. (November, 1948).

[7] R. Courant and K. O. Friedrichs, *Pure and Applied Mathematics Vol. 1, Supersonic Flow and Shock Waves* (Interscience Publishers, Inc., New York, 1948), p. 181.

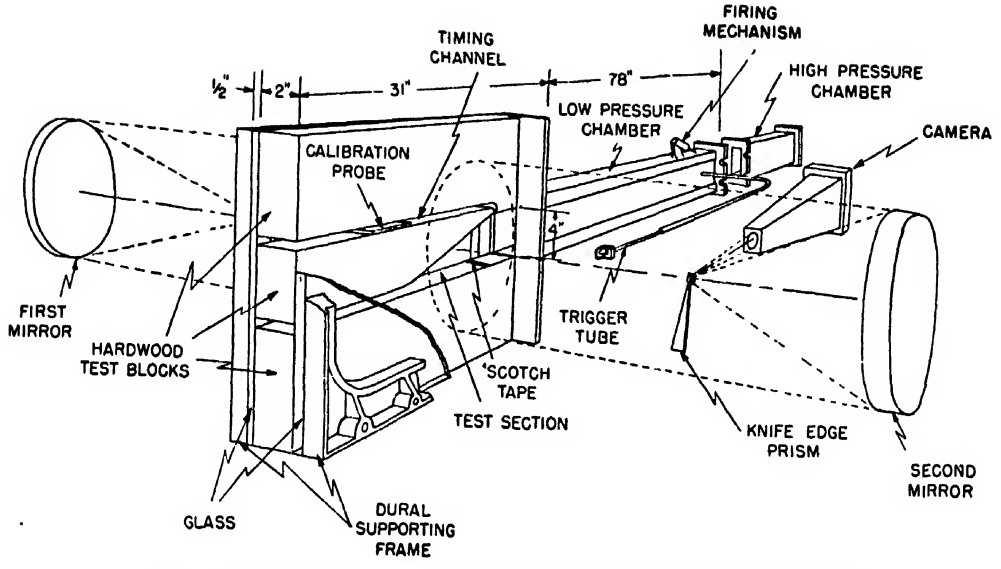


FIG. 1. Schematic drawing of shock tube and schlieren system used in this work.

were used to measure the shock strength and the time instant of the exposure.

APPARATUS

Standard Equipment

The shock tube used in these experiments (see Fig. 1) consisted of a rectangular steel tube 2×4 in. I.D. This tube was partitioned into high and low pressure chambers separated by a cellulose acetate diaphragm which could be pierced by a steel probe. By using various thicknesses of cellulose acetate up to 0.020 in., pressure differences between the chambers as large as 440 in. Hg could be used. In the calibration runs it was found that more consistent shock strengths were obtained if the diaphragm thickness was chosen to be close to its undisturbed bursting point when pierced. The end of the shock tube was equipped with a test section of interchangeable wooden blocks "sandwiched" between plate glass. A conventional schlieren apparatus was used to observe the flow. The spark source for the schlieren apparatus discharged 1.7 μf charged to about 9 kv. Exposures with a rotating mirror established the spark duration to be of the order of 1 μsec. A third trigger electrode was used which was energized via a thyratron.

Aerodynamic Instrumentation

The planned experimental program required the photography of shock waves by the schlieren system at various points during their propagation along the test section. Thus we needed to trigger the spark at controlled times and to measure the strength of the shock wave first for calibration purposes and later for the experimental work. The spark was triggered by a pressure wave which traveled down a small trigger tube (about $\frac{3}{16}$ in. I.D.—see Fig. 1). This trigger tube was connected to an orifice (about 2 in. from the diaphragm) and so received a pulse of high pressure air as the shock passed the orifice. Upon reaching the end of the trigger tube the pressure wave deflected a thin metal diaphragm a few thousandths of an inch to produce an electrical contact. This electrical contact was connected to the grid of the thyratron so as to trigger the spark. The trigger tube actually consisted of two metal tubes telescoping inside each other so that the spark could be triggered with appropriate lags to photograph the shock wave at various points along the test section with a reproducibility of about ±10 μsec. Once one photograph of the shock wave was obtained for a given shock strength, pictures

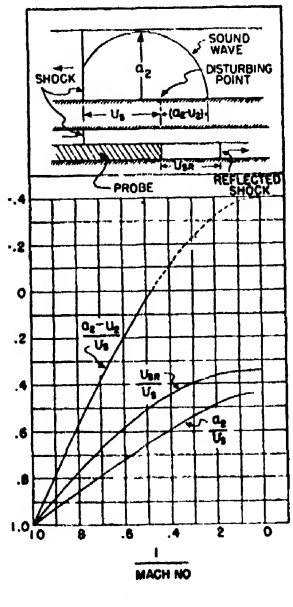


FIG. 2. Velocity ratios used for photographically measured shock strength by the application of weak waves (assumed infinitesimal in these calculations). The velocity ratio of a rearward moving weak wave $(a_1-u_1)/u_s$ was used where the shock strength was high and the velocity ratio of a normally reflected shock was used for measuring the strength of weak shocks, e.g., the timing wave.

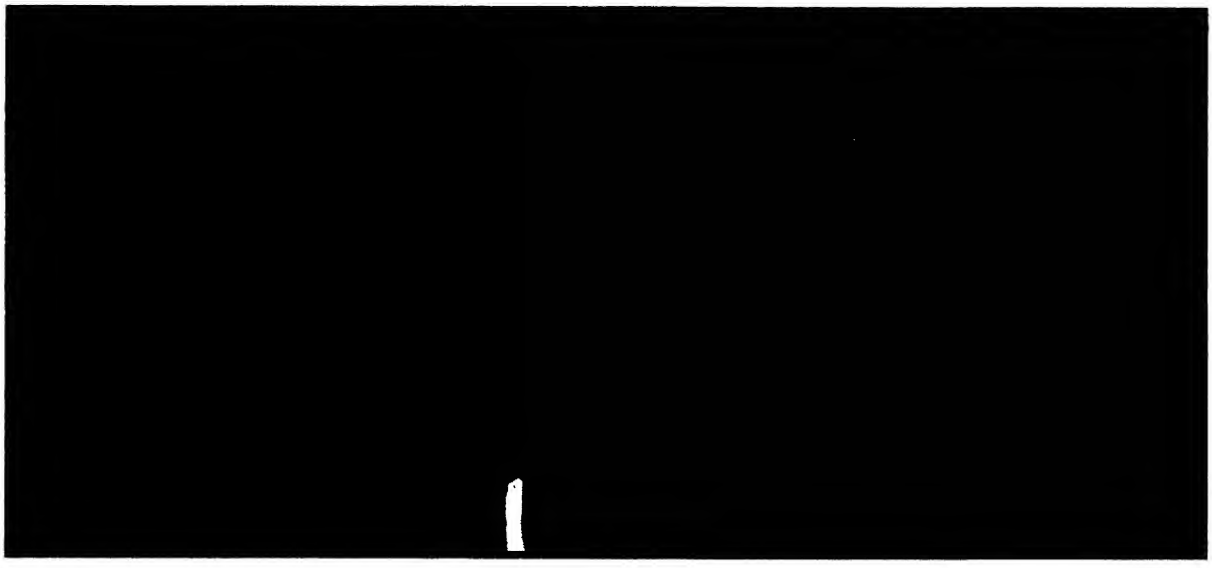


FIG. 3. A schlieren photograph of reflections from a single thickness of Scotch Tape. The left edge of the vertical marker corresponds to the leading edge of the (invisible) Scotch Tape.

at other points in the test section could be obtained by extending or retracting the trigger tube approximately (within five percent) the distance between the new and old shock positions.

A more accurate method of recording the actual instant of exposure was through the use of the timing wave. In Fig. 1 it can be seen that a passage about ½ in. high is included above the test section. This was connected to the main chamber by a narrow slit (0.015 in. wide and one-third the width of the channel). When the main shock passed this slit a very weak wave propagated along this upper channel at slightly greater than the speed of sound. The actual strength of this wave was measured with a calibration probe as will be discussed later. Its position at the instant of exposure was recorded along with the main shock wave on the schlieren photograph. Thus since the velocity of this wave is known from the calibration probe measurement (see below), its position is a measure of the time instant of exposure.

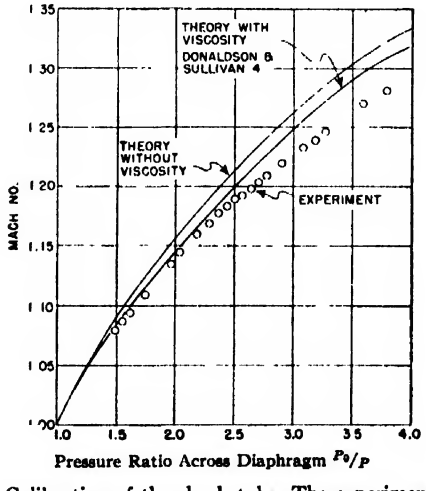


FIG. 4. Calibration of the shock tube. The experimental shock Mach numbers were obtained by the Scotch Tape technique (Fig. 3).

Shock Strength Measurement

Shock strengths were measured in all cases by the comparison of the speed of a reflected wave with the speed of a shock. The simplest technique is to place a piece of ordinary Scotch Tape on the floor of the shock tube in such a way that the passage of a normal shock will produce two weak cylindrical waves. The position of the upstream intersection of the rearward moving cylindrical wave with the wall and the position of a normal shock wave were measured relative to one edge of the Scotch Tape as shown in Fig. 2. The ratio of the distances of these waves from the disturbing point is identical with the ratio of the wave velocities. A photograph of the reflections from a piece of Scotch Tape is shown in Fig. 3. To measure the departure of these reflected waves from sound waves several different thicknesses of Scotch Tape were added to the original to produce stronger reflections. The apparent Mach numbers were plotted against tape thickness and the results extrapolated to zero thickness. No correction for a single thickness of Scotch Tape was required within the experimental scatter of the circles of the calibration on Fig. 4.

In order to calculate the strength of the shock from the ratio of the velocity of the reflected wave to the velocity of the normal shock wave, we need to know the strength of the shock wave in terms of u_2, a_2, and U_S where u_2 is the fluid velocity in the direction of the shock wave motion, a_2 is the velocity of sound behind the

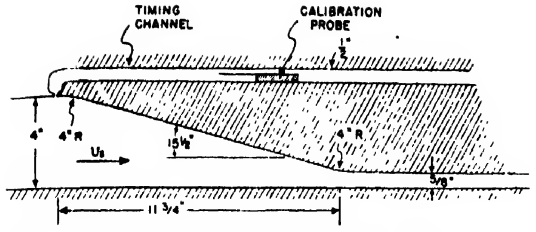


FIG. 5. Converging channel used in experimental work. The timing channel used to determine the exact time of the exposure is also illustrated.

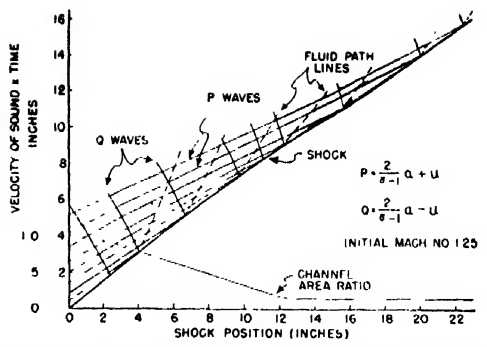


FIG. 6. Typical numerical integration along characteristics of the propagation of a shock wave (followed by no other disturbances) in a converging channel.

shock wave, and U_S is the velocity of the shock. It is desirable to transform to coordinates moving with the shock wave. Using this coordinate system the standard tabulations of the Rankine-Hugoniot relations can be used to give a relation between the shock Mach number and $(a_2-u_2)/U_S$. This relationship is plotted in Fig. 2. When the flow behind the shock becomes supersonic (the dashed curve on Fig. 2), this ratio would probably not be useful as simply measuring the Mach angle of the steady waves from the Scotch Tape. Similarly the wave moving across the channel can be used. In this case we measure the ratio a_2/U_S. This ratio is also plotted in Fig. 2. It can be seen that this ratio does not vary as rapidly with Mach number as the ratio between the upstream moving wave and the shock velocity. Therefore it does not lead to shock strength results as precise as those obtained from the upstream moving wave.

Normal reflection of the shock was used when the Scotch Tape reflection from a weak shock was no longer visible. A probe which is diagrammed in Fig. 2 was employed. In the lower part of this probe the wave was reflected normally whereas in the upper part it proceeded

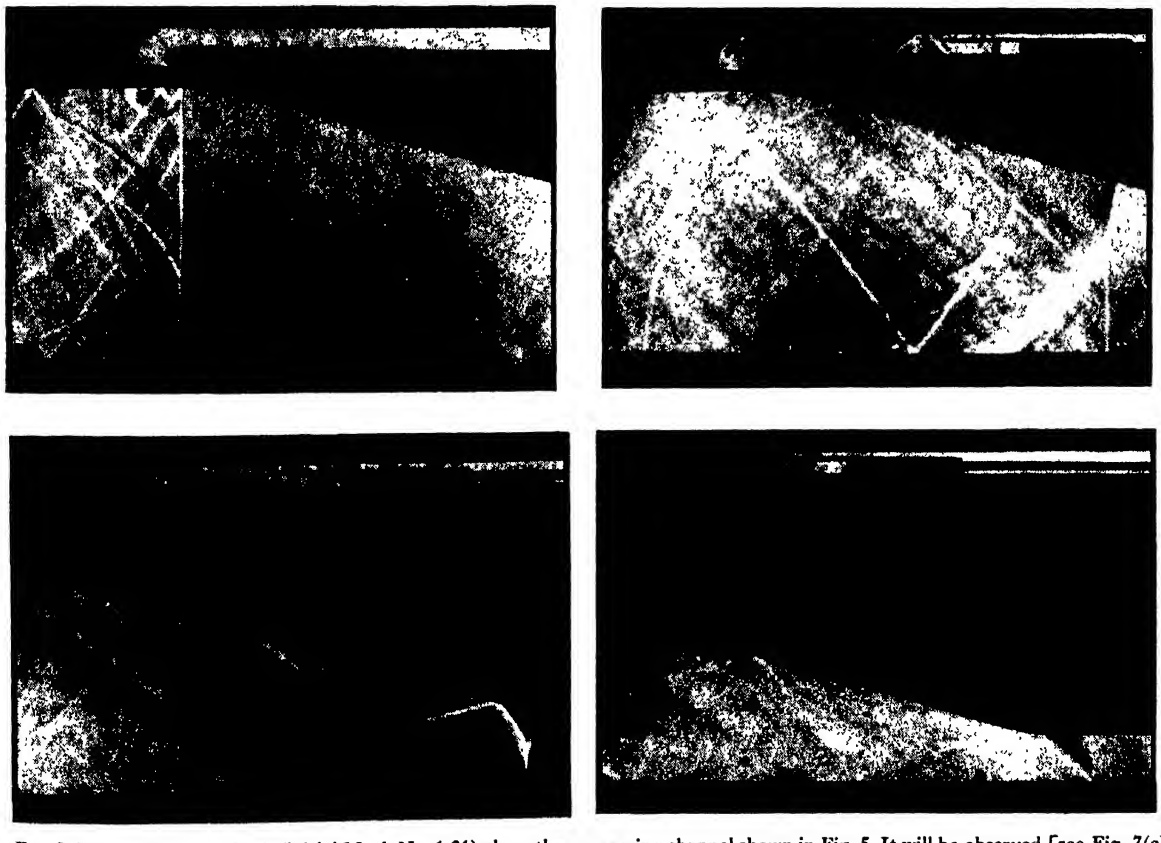


FIG. 7. Propagation of a shock (initial Mach No. 1.21) along the converging channel shown in Fig. 5. It will be observed [see Fig. 7(c)] that the shock rapidly assumes a cylindrical shape and then after a period of adaptation to the converging channel the one-dimensional theory of characteristics should apply.

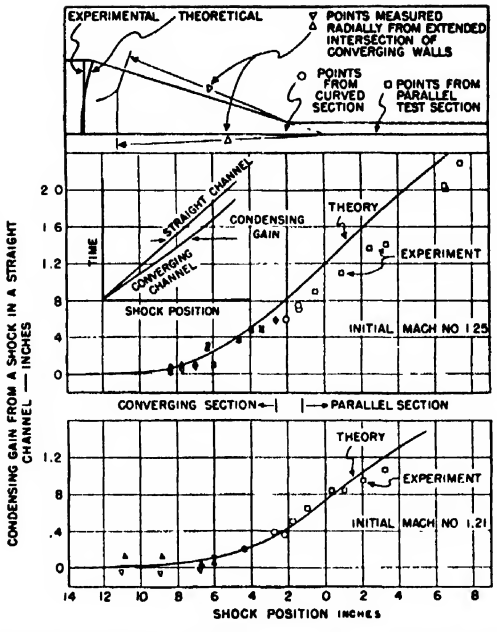


FIG. 8. Comparison of results obtained by numerical integration by the method of characteristics with experimental photographs such as Fig. 7.

with only slight disturbance. The ratio of the speed of the reflected wave to the speed of the initial wave can be used in a manner similar to the above for the determination of the shock strength. This ratio is also plotted in Fig. 2. The normal reflection technique was useful in finding the strength of the timing wave at various points along the timing channel, which had a Mach number of about 1.01.

Calibration of the Shock Tube

Before the shock tube can be conveniently used, a relation must be experimentally determined between the pressure ratio across the diaphragm and the Mach number of the resulting shock wave. This relationship can be determined from simple non-viscous theory by standard methods.[7] A viscous correction has recently been computed by Donaldson and Sullivan[4] which allows for the weakening of the shock wave by the boundary layer formed behind the shock wave. Both of these theoretical calibrations are plotted on Fig. 4. A series of photographs (such as Fig. 3) was taken at various pressure ratios using the Scotch Tape technique for measuring the shock strength. The measured shock Mach number is plotted against the pressure ratio on Fig. 4. It is seen that the Mach number obtained is always somewhat lower than the theory indicates even including wall boundary layer effects. This departure from the theoretical calculation including viscous effects might be expected in view of possible obstructions remaining at the diaphragm and miscellaneous losses due to partial reflections due to roughness in the shock tube and the test section.

APPLICATION TO THE PRELIMINARY STUDIES OF CONDENSING SHOCK WAVES

As was mentioned in the Introduction, our first experiments with the shock tube were directed to the study of the limitations of the one-dimensional method of characteristics for unsteady flows. It is assumed in this theory that conditions are uniform across the channel. On the other hand, if changes in the channel area occur, the effects will be felt first near the wall and only later in the center of the channel. In fact, it is clear from the photographs taken in the test channel that the effects of the wall led to a reflected wave similar to a triple shock intersection. This intersection point divides the shock into two portions, one of which has been affected by the change of wall contour and the other of which has not been affected by this change. We were also interested in determining whether a wave in a converging passage would ultimately assume a cylindrical form, i.e., to see whether the cylindrical form is stable for a condensing shock.

The channel chosen for these experiments is illustrated schematically in Fig. 5. The one-dimensional characteristic theory for this channel was computed by the methods of reference 6, and one such computation is shown in Fig. 6. No viscous effects are included in this calculation and no attempt to correct for departures from one-dimensional flow has been made.

The boundary conditions for the calculation were, first, that the flow is initially at rest with sonic velocity[6] a_o ($P=Q=[2/(\gamma-1)]a_0$) and, second, that no later waves propagated into the test section following the shock (i.e., at the beginning of the convergent section $P=[2/(\gamma-1)]a+u$ is a constant determined from the known initial shock strength).

A series of photographs was taken for two different shock Mach numbers (1.21 and 1.25) and typical photographs from the series for $M=1.21$ are shown in Figs. 7(a)–7(d). The earlier figures can be best described by saying that the effect of the wall deflection was propagated as a triple shock intersection back and forth across the channel. The later figures show that the shock does tend to approach a cylindrical form and thus indicate that the cylindrical form of the shock wave is stable in converging channels for the conditions of these experiments.

A comparison is given in Fig. 8 between the position of the shock wave as determined experimentally and the positions calculated by the method of characteristics. It will be noticed that the two ends of the shock approach the same radial position in about the time the triple shock intersection crosses the channel once.

The Magnetostriction of Permanent Magnet Alloys

E. A. NESBITT
Bell Telephone Laboratories, Murray Hill, New Jersey
(Received April 24, 1950)

In order to obtain a better understanding of the mechanism of coercive force in modern permanent magnets, magnetostriction measurements have been made on various alloys having coercive forces from 50 to 600 oersteds. The results can be summarized by discussing two types of alloys. First are the older carbon-hardening permanent magnets, and for these alloys high coercive force and high magnetostriction occur together. Second are the newer carbon-free permanent magnets and for these alloys high coercive force does not occur with high magnetostriction. In fact for the Mishima alloys having compositions near 29 percent nickel, 12.5 percent aluminum, and 58.5 percent iron, cooled at the rate of 3°C per second (coercive force 400 oersteds), the magnetostriction actually passes through zero. This is contrary to the classical strain theory of coercive force which states that the latter is proportional to the product of the magnetostriction and internal stress. To explain the mechanism of coercive force for these alloys it is necessary to resort to more recent theories.

INTRODUCTION

THE main purpose of this work has been to measure the magnetostriction of permanent magnet alloys having coercive forces from 50 to 600 oersteds, with the hope that the results would aid in understanding the mechanism of coercive force. At the present time the data on magnetostriction in this field are limited, although data of this type on materials of low coercive force are voluminous and have been of great value in interpreting the properties of materials of high permeability.

According to the older theories of Kersten,[1] Becker,[2] Kondorsky,[3] and others, a coercive force occurs when the material has (1) large internal strains that vary from place to place (strain gradients) and (2) large magnetostriction at saturation (change in length from demagnetized state to saturation). This classical strain theory of coercive force is expressed as follows:

$$H_c \doteqdot \lambda_s \sigma_i / I_s$$

where λ_s is the magnetostriction at saturation, σ_i is the amplitude of internal stress and I_s the saturation intensity of magnetization. Because of theoretical and experimental work, the validity of this theory has been questioned.

Néel[4] has developed a theory of coercive force in which he considers fluctuations in the magnetization of the elementary domains. Fluctuations in the *directions* of magnetization are caused by elastic disturbances of the lattice, and fluctuations in *intensity* of magnetization are caused by the presence of inclusions or cavities. Because of these fluctuations there will be magnetic poles throughout the material and associated with them will be energy, the distribution of which will largely determine the positions of the domain walls and their movement with a field.

Stoner and Wohlfarth[5] consider many materials having high coercive force (*magnets*) to be made basically of fine particles (this term including both actual powder grains and atomic segregates or islands in alloys) which act as single magnetic domains. The coercive force of these domains depends on one or more of three types of anisotropy, namely, strain, crystal and shape anisotropy. Strain anisotropy is a function of the strain and the magnetostriction. Crystal anisotropy depends on the crystal structure, and is measured by the anisotropy constant. Shape anisotropy depends upon the shape of the particle in such a way that it is zero for

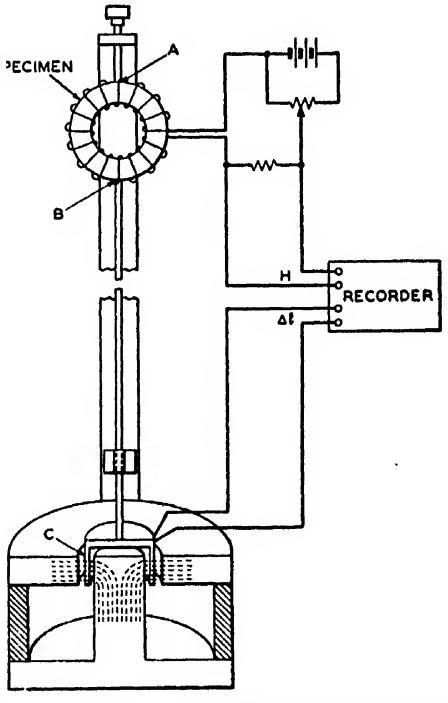


FIG. 1. Schematic diagram of magnetostriction instrument. Fluxmeter may be used in place of recorder.

[1] M. Kersten, Zeits. Tech. Physik 12, 667–9 (1931).
[2] R. Becker, Physik. Zeits 33, 905–13 (1932).
[3] E. Kondorsky, Physik. Zeits Sowjetunion (U.S.S.R.) 8, 153 (1937).
[4] L. Néel, Comptes Rendus 224, 1488 (1947).
[5] Stoner and Wohlfarth, Phil. Trans. Roy. Soc. (London) 240, 599–644 (1948).

FIG. 2. Photograph of specimen, box, and winding details.

a material having spherical particles and high for a material composed of long, thin needles. Magnets composed of fine particles of this kind, and having high coercive force, have been produced by Dean and Davis[6] (powdered iron, cobalt, etc.), Guillaud[7] (powdered MnBi, MnAs, etc.), Weil,[8] and others. The theory has been discussed by Néel, and Stoner and Wohlfarth and by Kittel.[9] In the case of iron, the particles must be of the order of 2×10^{-6} cm or smaller in order to act as single magnetic domains.

The measurements here reported make it necessary in the case of some important alloys to resort to the new theories of Néel, and Stoner and Wohlfarth and to reject the older strain theory. For example in some of the Mishima alloys zero magnetostriction was obtained with an accompanying high coercive force and this is incompatible with the strain theory. In other materials such as high cobalt steel the strain theory may still apply.

FIG. 3. Photograph of cylindrical coil, bearings, and circular gap in magnet.

[6] Dean and Davis, U. S. Patent 2,239,144.
[7] Guillaud, Thesis, Strasbourg (1943).
[8] L. Weil, Comptes Rendus 225, 229 (1947).
[9] C. Kittel, Phys. Rev. 70, 965 (1946).

APPARATUS

The first phase of this work was the development of a new method for measuring magnetostriction which was originally devised by H. J. Williams. The method is a magnetic one and has the desirable feature of recording automatically the change in length of a specimen as a function of the magnetizing force. Figure 1 shows a schematic diagram of this instrument.

The toroidal specimen is rigidly clamped to the brass frame at the point A as shown in Fig. 1. At a point

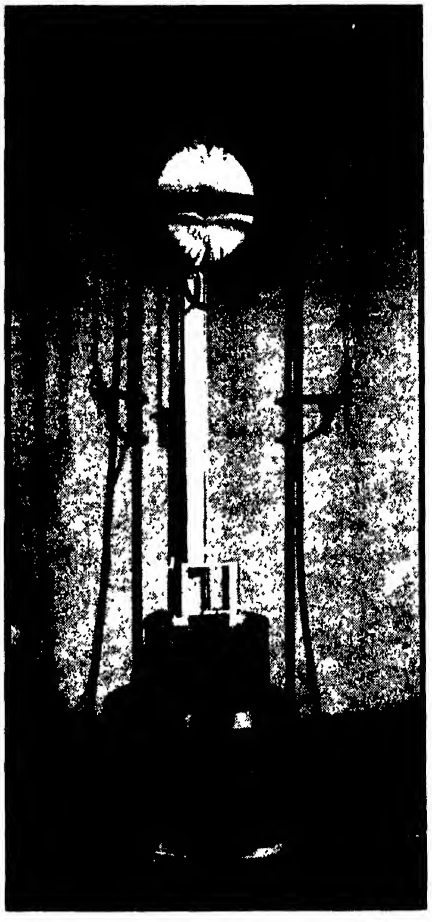


FIG. 4. Photograph of complete assembly of magnetostriction instrument.

diametrically opposite, a brass rod is soldered to the specimen and this rod supports a cylindrical coil C which is suspended in a uniform radial magnetic field supplied by permanent magnets. Any vertical displacement of the cylindrical coil causes a change in the flux linkages of the coil and this is measured by the recorder. The recorder is a Cioffi recording fluxmeter which consists of two elements, one drives a paper drum proportional to the magnetizing force, the other drives the pen across the paper in proportion to the change in flux linkages. In practice the instrument is actuated by

sending current through the magnetizing winding of the specimen. This causes two things to happen. First, sending current through the magnetizing winding also sends current through the recorder and this causes the paper drum to turn by an amount proportional to the magnetizing force. Second, sending current through the magnetizing winding magnetizes the specimen and changes its diameter thus causing a vertical displacement of the cylindrical coil and as a result the pen then moves an amount proportional to the change in flux linkages. Both of these actions occur simultaneously, thus a curve is traced of the change in length of a specimen *versus* magnetizing force.

Figure 2 shows the specimen, the box in which the specimen is suspended but not allowed to touch, the cover for the box and the completely wound specimen. The specimen shown is approximately 4 in. I.D., 4.6 in. O.D., and is either a cast or a machined ring. Specimens 2 in. I.D., 2.5 in. O.D. were also used. The brass extensions are soldered to the specimen, the heavy one fastens to the brass frame, the light one supports the

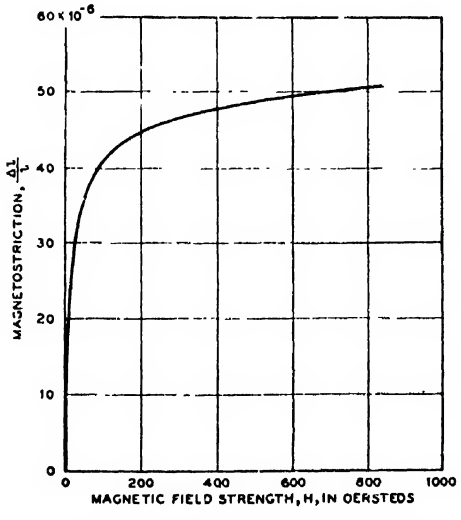


FIG. 5. Longitudinal magnetostriction of Vanadium-Permendur.

cylindrical coil. The box and its cover are watercooled; this is important because it is necessary to keep the temperature of the specimen constant during the measurement. The specimen shown in Fig. 2 is wound with 1850 turns of No. 18 Formex wire. The smaller specimens were wound with approximately 800 turns of No. 19 enameled silk copper wire. It is important to make the windings uniform and continuous in order to avoid magnetic pick-up between the specimen and the cylindrical coil. In some cases it is desirable to have a cylindrical magnetic shield surrounding the toroidal magnetizing winding.

Figure 3 shows the cylindrical coil and its associated bearings for guiding the motion of the coil. The coil is wound with 1600 turns of No. 38 enameled copper wire. The bearings are fitted to have a minimum of frictional

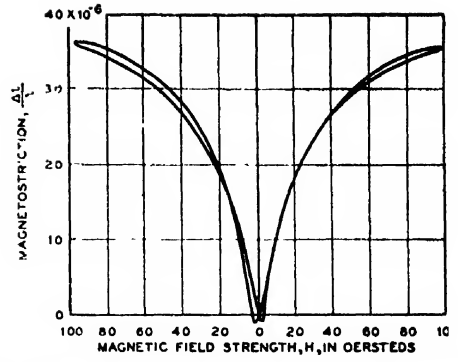


FIG. 6 Magnetostrictive loop of Vanadium-Permendur.

resistance and yet prevent side play of the coil *C*. The coil is placed in the circular gap.

Figure 4 shows the complete assembly of the instrument. The instrument is mounted on a base which is supported by a Julius suspension, part of which can be seen in the photograph. The base is oil damped and the entire instrument enclosed in an air shield (not shown in photograph) since it is important to reduce all mechanical and wind vibrations to a minimum. The rubber tubes shown in the photograph carry water to cool the box containing the specimen. The distance between the specimen and the coil should be at least two feet.

In order to test the accuracy of this new instrument several of the soft magnetic materials were measured since these are freer from cracks, easier to magnetize and chemically more homogeneous. The latter factor is important in measuring magnetostriction because any inhomogeneity in the specimen causes a pole to occur and this results in an undesirable extraneous force between the specimen and the magnetizing coil. Figure 5

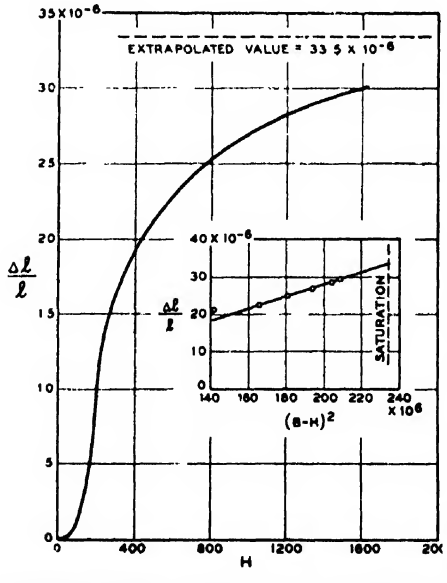


FIG. 7. Magnetostriction of Alnico V after cooling from 1300°C at 2°C per second.

TABLE I. Magnetostriction of Alnico V.

Treatment	Size	$\frac{\Delta l}{l}10^6$ $H=1600$	$\frac{\Delta l}{l}10^6$ maximum	Magnetic properties B_r	H_c	B_m	H for B_m
1300°C—cooled approximately 2°C per second	4 in. I.D. 4.5 in. O.D.	30	33.5	11650	270	16300	1510
1300°C—cooled approximately 2°C per second		27.5	35	10900	140	16420	1920
1300°C—cooled approximately 2°C per second plus 585°C for 8 hr.	2 in. I.D. 2.5 in. O.D.	21.8	22.3	9200	450	15150	2180
1300°C—cooled approximately 2°C per second to 820°C then oil quenched		26	26	9060	13	15300	1120
Sand casting	4 in. I.D. 4.5 in. O.D.	19	22.3	9620	201	15100	1940
1000°C—1 hr.–furnace cooled	2 in. I.D. 2.5 in. O.D.	17.2	25	8550	298	14900	1930
1300°C—cooled approximately 2°C per second in a transverse magnetic field	2 in. I.D. 2.5 in. O.D.	36	43	5140	154	15220	1900
1300°C—cooled approximately 2°C per second in a longitudinal magnetic field	2 in. I.D. 2.5 in. O.D.	3.5		12900	252	16200	1460
Cooled in longitudinal field, quenched at 820°C	2 in. I.D. 2.5 in. O.D.	2.5	2.5	13250	12.8	14500	147
Cooled in transverse field, quenched at 820°C	2 in. I.D. 2.5 in. O.D.	38	38	610	17.2	15400	895

TABLE II. Magnetostriction of alloys in the vicinity of the Alnico V composition.

Composition Ni	Al	Co	Cu	Fe	Treatment	Ring size	$\frac{\Delta l}{l}10^6$ $H=1600$	$\frac{\Delta l}{l}10^6$ maximum	B_r	H_c	B_m	H for B_m
14	2	24	3	57	1000°C—1 hr.–F.C.	2.0 in. I.D. 2.5 in. O.D.	30.4	38	5850	18	21600	1160
14	6	24	3	53	1300°C—cooled 2°C per second to 820°C then oil quenched		19.2	19.2	4010	22.4	17100	1120
14	6	24	3	53	Sand cast	4.0 in. I.D. 4.5 in. O.D.	25.5	25.6	4540	56.6	17000	1700
14	6	24	3	53	1300°C—cooled approx. 2°C per second	4.0 in. I.D. 4.5 in. O.D.	25.3	25.3	7120	86	17000	1640
14	6	24	3	53	1000°C—1 hr.–F.C.	2.0 in. I.D. 2.5 in. O.D.	14.2	14.4	6990	103	17100	1570
14	10	24	3	49	Sand cast	1.9 in. I.D. 2.5 in. O.D.		19.0	9330	4.0	15500	1590
14	10	24	3	49	1000°C—1 hr.–F.C.	2.0 in. I.D. 2.5 in. O.D.	14.2		7570	249	16300	1900

TABLE III. Magnetostriction of alloys of varying coercivity.

Composition	Treatment	Ring size I.D.	$\frac{\Delta l}{l}\times 10^6$ $H=1600$	$\frac{\Delta l}{l}\times 10^6$ maximum	B_r	H_c	B_m	H for B_m
Tool steel (0.9 percent carbon)	800°C—5 min.—brine quench +100°C -48 hr.	2 in.	−2.8		10850	45	19100	1250
Low cobalt steel (0.9C—4.75Cr—1.25W 8.5Co—bal. Fe)	900°C—5 min.—oil quench +100°C—24 hr.	2 in.	+14.5	+14.5	9000	93	16720	1340
Honda steel (0.8C—4W—2Cr—36Co—bal. Fe)	940°C—7 min.—oil quench +100°C—24 hr.	2 in.	+33.5	+40.5	10180	183	19900	1860
Remalloy (71Fe—12 percent Co—17Mo)	1300°C—20 min.—oil quench	2 in.		24	4270	2.6	17600	354
Remalloy (71Fe—12 percent Co—17Mo)	1300°C—20 min.—oil quench +685°C—1.5 hr.	4 in.	+11.8	+17	9580	305	16400	1860
Vicalloy (37.5Fe—52Co—10.5V)	1000°C—30 min. -brine quench	2 in.	+34.2	+47.5	4900	77	16300	2100
Vicalloy (37.5Fe—52Co—10.5V)	1000°C—30 min.—brine quench +600°C—1 hr.	2 in.	+26.0	+34	7500	320	14100	1660
Alnico V	1300°C—20 min.—cooled 2°C per second	4 in.	+30	+33.5	11650	270	16300	1510
Alnico V	1300°C—20 min.—cooled 2°C per second	4 in.	+27.5	+35	10900	140	16420	1920
Alnico V	1300°C—20 min.—cooled 2°C per second +585°C—8 hr.	2 in.	+21.8	+22.3	9200	450	15150	2180
Cunico (21Ni—50Cu—29Co)	Commercial heat treatment	2 in.	+3.7	+7.5	4140	445	8820	2400
Vicalloy tape (38Fe—52Co—10V)	800°C—1 min.—H_2 quench	2 in.	+10.5	+11	1700	195	7150	1900
23Ni—12Al—65Fe	1300°C—30 min.—cooled 5°C per second	2 in.		−1.0	6800	356	12800	1960
29Ni—13Al—58Fe	1300°C—30 min.—cooled 5°C per second	2 in.	−3.0	−6.2	5000	482	10880	1965
29Ni—13Al—58Fe	1300°C—30 min.—cooled 5°C per second +650°—2 hr.	2 in.	−5.5	−10.5	5080	461	10850	1730
29Ni—13Al—58Fe	As cast in sand	2 in.	−5.9	−15.0	4850	560	11320	2040
29Ni—13Al—58Fe	1000°C—1 hr.-F.C.	2 in.	−7.0	−17.5	2700	134	12200	2400
29Ni—10Al—63Fe	1350°C—cooled 4°C per second	2 in.	+7.0	+13	5050	340	10530	1720
29Ni—12Al—61Fe	1350°C—cooled 4°C per second	2 in.	+3.6	+5.5	4640	420	10700	1945
29Ni—12.5Al—60.5Fe	1350°C—cooled 4°C per second	2 in.	−2.0	−4.5	4400	400	11180	2410

shows an actual magnetostrictive trace obtained with this instrument in conjunction with the Cioffi recording fluxmeter, using a specimen of Vanadium-Permendur, an alloy of 49 percent iron, 49 percent cobalt, and 2 percent vanadium. At a field strength of 830 oersteds, a magnetostriction of 51×10^{-6} was obtained. The specimen was approximately 4 in. I.D., 4.5 in. O.D., and had been annealed at 1000°C for 1 hour and slowly cooled. This agrees closely with values obtained by other investigators, for example, on this alloy not containing the vanadium, Honda[10] obtained a magnetostriction of approximately 50×10^{-6} with $H=500$ oersteds. Sometimes it is desirable to take magnetostrictive loops over limited field strengths and Fig. 6 shows that this

[10] K. Honda, *Magnetic Properties of Matter* (Tokyo, 1928), p. 111.

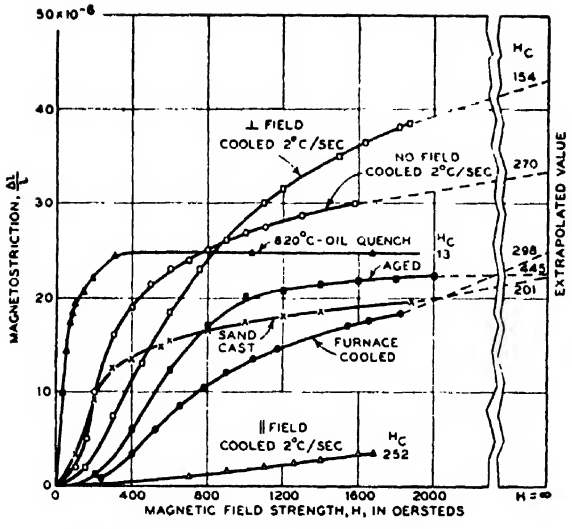


FIG. 8. Magnetostriction of Alnico V for seven different treatments.

instrument is capable of taking symmetrical loops of this type. The specimen is the same as the one used for Fig. 5. The specimen was first magnetized at its remnant value and the magnetizing force then applied in the opposite direction. The magnetostriction dipped slightly then rose to a maximum value of 35.6×10^{-6}, and then as the H was decreased, the magnetostriction decreased to zero. The H was reversed, the magnetostriction dipped again and then rose to a maximum of 36.2×10^{-6} and then as the H was decreased again, the magnetostriction decreased to zero, its starting point.

In many of the experimental curves, the saturation magnetostriction was not reached since magnetizing forces of only about 2000 oersteds were used. It is difficult to saturate completely permanent magnets in their hardened state, and this is especially true when toroidal specimens are used. The use of toroidal specimens has several advantages, however, because by this method one avoids as much as possible the translational forces caused by non-uniform fields and also avoids errors resulting from the forces between free poles. Figure 7 shows trace of the magnetostriction as a function of the magnetizing force for Alnico V and an extrapolated curve (data obtained from trace) of magnetostriction as a function of $(B-H)^2$ for obtaining the saturation magnetostriction. Sometimes it was also necessary to plot $(B-H)$ *vs.* $1/H$ to obtain the saturation magnetization by extrapolation.

The measurement of zero magnetostriction on a polycrystalline specimen does not necessarily mean that the magnetostrictions in the various crystallographic directions are zero. It is possible that a negative magnetostriction in the [111] direction could be balanced somewhat by two positive magnetostrictions in the [100] and [110] directions as is the case in iron. However, an exact balance of these crystallographic magnetostrictions for all values of flux density is unlikely. When the polycrystalline magnetostriction is found to be low at all inductions, it is believed that the magnetostriction is low in all crystallographic directions.

RESULTS

The data obtained may be divided into four categories.

(1) The magnetostriction of Alnico V and its variation with field and heat treatment.

(2) The magnetostriction of alloys having compositions near that of Alnico V.

(3) The magnetostriction of permanent magnet alloys of varying coercivity. The purpose of these experiments is to note any general relationship between magnetostriction and coercive force.

(4) The magnetostriction of several alloys of special interest which cannot be classified as permanent magnets. A specimen of Ferroxcube received from the American Philips Company, through the kindness of Dr. F. Brockman, and specimens of oriented iron-cobalt were measured. An iron-cobalt specimen yielded a magnetostriction of 130×10^{-6} one of the highest values ever reported.

The data for the first three parts are summarized in Tables I–III and they will now be discussed in detail.

Alnico V is probably the most important permanent magnet material. This alloy and associated alloys which are modifications of Alnico V are the only commercial permanent magnets which respond to heat treatment in

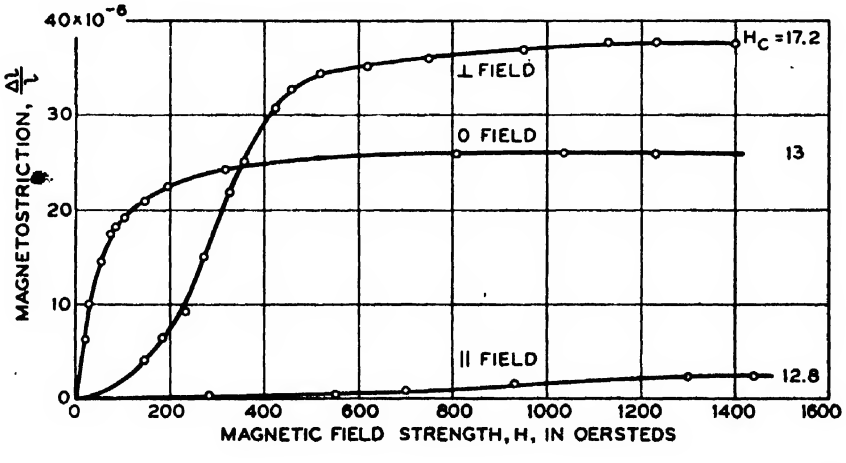


FIG. 9. Magnetostriction of Alnico V quenched from 820°C.

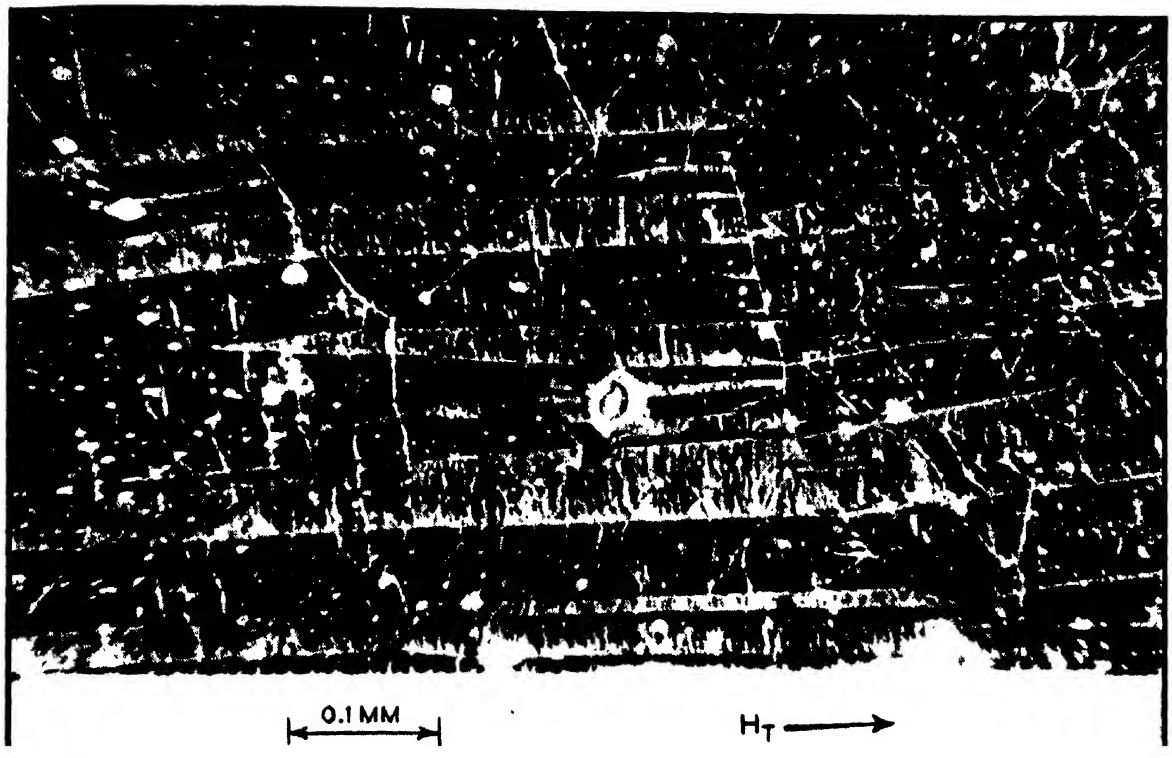


FIG. 10. Powder pattern photograph of Alnico V heat treated in a longitudinal magnetic field, the direction of which is shown by the arrow.

a magnetic field. Suggestions[11] have been made for explaining its response to heat treatment in a field and the nature of its hardening mechanism, and these suggestions have been aided by measurements of magnetostriction.

Figure 8 shows the magnetostriction of Alnico V as a function of magnetizing force for seven different treatments. The saturation magnetostriction values of the quenched, sandcast, and furnace cooled states are respectively 26×10^{-6}, 22.3×10^{-6}, and 25×10^{-6}. In the quenched specimen the precipitation process has not yet begun, in the sandcast specimen the process has not yet advanced to the point for optimum magnet properties while in the furnace-cooled specimen the process has advanced beyond the optimum point. The optimum cooling rate for Alnico V is approximately 2°C per second from 1300°C and the curve for zero field shows the magnetostrictive properties obtained on a specimen cooled at this rate. At an H of 1600 oersteds, the magnetostriction was 30×10^{-6} and the extrapolated value for saturation was 33.5×10^{-6}. The magnetic properties of this specimen are residual induction (B_r) of 11650 gauss, coercive force (H_c) of 270 oersteds, and the B for an H of 1510 oersteds is 16300 gauss. Another specimen of Alnico V (not shown in figure) heat treated in a similar manner had a magnetostriction of 27.5×10^{-6} at an H of 1600 oersteds and an extrapolated value of 35×10^{-6}. These two specimens cooled at the optimum rate yielded the highest magnetostriction for specimens of Alnico V not heat treated in a magnetic field. When Alnico V is completely hardened the saturation magnetostriction is lowered, a value of 21.8×10^{-6} was obtained at an H of 1600 oersteds the extrapolated value was 22.3×10^{-6} as shown in the curve for the aged specimen. The magnetic properties of this specimen were $B_r=9200$ gauss, $H_c=450$ oersteds and a saturation magnetization of 14150 gauss. The results given so far are for Alnico V not heat treated in a magnetic field and a random domain distribution was believed to exist. Specimens of Alnico V heat treated in longitudinal and transverse magnetic fields were also measured. The magnetostriction of the specimen heat treated in a longitudinal field was low as would be expected from the domain theory. A value of only 3.5×10^{-6} at an H of 1680 oersteds was obtained as shown in the curve for the parallel magnetic field. The magnetic properties of this specimen were $B_r=12900$ gauss, $H_c=252$ oersteds and $B=16200$ gauss for an H of 1460 oersteds. The magnetostriction of the specimen heat treated in a transverse field was high as would be expected from domain theory. A value of 36×10^{-6} was obtained at an H of 1600 oersteds and the extrapolated value was 43×10^{-6}. This is higher than that obtained on the specimens cooled at the same rate but with no applied field during the heat treatment. The magnetic properties of the

[11] Kittel, Nesbitt, Shockley, Phys. Rev. 77, 839–840 (1950).

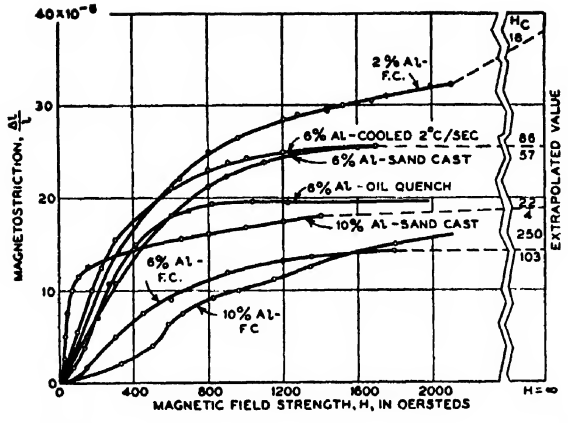


FIG. 11. Magnetostriction of several alloys having compositions near Alnico V. F. C. indicates furnace cooled.

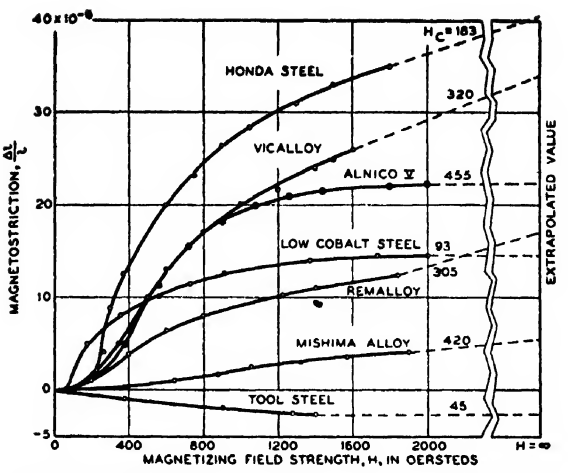


FIG. 12. Magnetostriction of permanent magnet alloys of different coercivity.

specimen were $B_r = 5140$ gauss, $H_c = 154$ oersteds, and $B = 15220$ gauss for an H of 1900 oersteds.

Figure 9 compares the magnetostriction of three specimens of Alnico V which were quenched from 820°C. The first specimen was cooled in a transverse field and its magnetostriction was 38×10^{-6}. This value is almost one and one-half times higher than the magnetostriction of the second specimen which was cooled in zero field. This is in exact accordance with domain theory which predicts that a specimen having the domains oriented transversely will have one and one-half times the magnetostriction of a specimen having random domain orientation. The third specimen was heat treated in a longitudinal field and its magnetostriction was 2.5×10^{-6}. This low value is obtained because the longitudinal field orients the domains either parallel or antiparallel to the direction of the applied field and so very little change in length is observed upon magnetization. A powder pattern photograph of the domains in this ring is shown in Fig. 10, taken with H. J. Williams. These large domains or domain aggregates extend around the circumference of the ring in the same direction that the field was applied during heat treatment. As a result of this phase of the investigation, we may say that Alnico V normally has a high magnetostriction and that it varies greatly with heat treatment from 2.5×10^{-6} to 43×10^{-6}. The behavior of the specimens heat treated in a magnetic field conforms with modern domain theory.

The second phase of this work consisted of measuring the magnetostriction of several alloys having aluminum contents in the vicinity of standard Alnico V (8 percent aluminum) to note how the magnetostriction varies with composition. The sum of the aluminum and iron percentages was kept constant. Figure 11 shows seven curves of magnetostriction as a function of magnetizing force for alloys of two, six, and ten percent aluminum. The magnetostriction of the six percent aluminum alloy for sandcast, cooled 2°C per second, quenched, and furnace cooled states, are respectively 25.5×10^{-6}, 25.3×10^{-6}, 19.2×10^{-6}, and 14.2×10^{-6} for an H of 1600 oersteds. With this alloy as with the standard 8 percent aluminum, the magnetostriction depends on heat treatment. Jellinghaus[12] mentioned that alloys containing 5 to 7 percent aluminum have a high Curie point but for some unknown reason fail to respond to heat treatment in a magnetic field. We now know that the 6 percent aluminum alloy also has a substantial value of magnetostriction although its value is only 75 percent that of the Alnico V. Alloys containing only 5 to 7 percent aluminum develop only a fraction of the coercive force of Alnico V. Magnetostriction values of the 10 percent aluminum alloy for the sand cast and furnace cooled states are respectively 17.5×10^{-6} and 12.9×10^{-6} for an H of 1400 oersteds. The furnace cooled specimen has an irregular curve indicating a heterogeneous structure. The magnetization curve has a similar irregularity showing that the magnetostriction is following the magnetization. The magnetostriction of the furnace cooled alloy with 2 percent aluminum was 30.4×10^{-6} at an H of 1600 oersteds and 38×10^{-6} at saturation. This alloy containing only 2 percent aluminum is not in the immediate vicinity of Alnico V and its magnetostriction is due primarily to the magnetic constituents of iron, nickel and cobalt and not to the aluminum.

The third phase of this work consisted of measuring the magnetostriction of a number of permanent magnets having different values of coercive force, as shown in Fig. 12. For tool steel at an H of 1600 oersteds the magnetostriction has a value of approximately -2.8×10^{-6}. This low negative value is also accompanied by a low value of coercive force (45 oersteds). The addition of cobalt to steel (along with several carbide producing elements such as tungsten and chromium) produces a series of permanent magnets of increasing coercivity. The magnetostriction of a low cobalt steel (8.5 percent cobalt) is shown in Fig. 12 and has a value of 14.3×10^{-6}

[12] W. Jellinghaus, Arch. Eisenhüttenw. 16, 247–251 (1943).

at $H=1600$ and a coercive force of 93 oersteds. The best of these cobalt steels is Honda steel and its magnetostrictive curve shown in Fig. 12 yields a value of 33.6×10^{-6} at an H of 1600 oersteds, accompanied by a coercive force of 183 oersteds. Thus we see that in the cobalt steel system for the alloys measured as the magnetostriction increases the coercive force does likewise. This result is consistent with the classical strain theory of coercive force (coercive force is proportional to the product of magnetostriction and stress) if we assume that the stress is substantially constant in these alloys. This seems like a reasonable assumption since the alloys are closely related to each other and their indentation hardness is approximately the same.

Remalloy,[13] a hot workable alloy of iron, cobalt, and molybdenum, and Vicalloy,[14] a hot or cold workable alloy of iron, cobalt, and vanadium, were measured. The magnetostriction of the Remalloy specimen as 11.6×10^{-6} at 1600 oersteds, and the extrapolated value was 17×10^{-6} (Fig. 12). The magnetostriction of the Vicalloy specimen was 26×10^{-6} at 1600 oersteds and the extrapolated value was 34×10^{-6}. The coercive force of the Remalloy was 305 oersteds, the coercive force of the Vicalloy was 320 oersteds while the respective magnetostrictions were 17 and 34×10^{-6}. In this instance, therefore, we have two different alloys with approximately the same coercive force but one alloy has twice the magnetostriction of the other. In addition the magnetostrictions of the Remalloy and Vicalloy specimens were measured in an unhardened state and the values obtained were respectively 24 and 47.5×10^{-6} (Table III). These values are considerably higher than those obtained on the hardened specimens, it being generally true that age-hardening lowers the magnetostriction of an alloy.

The magnetostriction curve for the hardened Alnico V (not heat treated in a field) is replotted in Fig. 12. The magnetostriction was 21.8×10^{-6} at 1600 oersteds and the extrapolated value was 22.3×10^{-6}, and the coercive force was 450 oersteds in the same specimen.

The magnetostriction of the Mishima alloy (29 percent Ni, 12 percent Al, 59 percent Fe), shown in Fig. 12, was so low for its accompanying coercive force of 420 oersteds that it was decided to investigate this system in detail. A series of Mishima alloys were made with nickel held constant at 29 percent and aluminum varying from 8 to 13 percent, and the results obtained are shown in Fig. 13. The magnetostriction of these alloys gradually diminishes with increasing aluminum content and actually passes through zero in the vicinity of 12 to 12.5 percent aluminum. According to the strain theory of coercive force, the coercive force should also pass through zero in this vicinity but the alloys in this range have coercive forces from 400 to 440 oersteds. Evidently the strain theory does not apply to these alloys and in all probability it does not apply to Alnico V since this alloy is directly related to the iron-nickel-aluminum system. The alloys were cooled from 1350°-1300°C at the rate of 2°-5°C per second and their magnetostriction depends to a certain extent upon this heat treatment. This is shown in Table III for the alloy containing 13 percent aluminum. Saturation magnetostriction values for the annealed, cast, aged, and cooled at 5°C per second states are respectively -17.5×10^{-6}, -15.0×10^{-6}, -10.5×10^{-6}, and -6.2×10^{-6}.

A number of the alloys having coercive forces from 50 to 600 oersteds are shown graphically in Fig. 14, in which coercive force is plotted against magnetostriction. In addition to the results shown in Figs. 12 and 13, magnetostriction is plotted for the alloy Cunico (50 percent Cu, 21 percent Ni, 29 percent Co), Vicalloy tape of low magnetic saturation, and two Mishima alloys. Information on the indentation hardness and magnetic saturation are also included. The indentation hardness probably bears some relation to the amplitude of the internal stress. Also in a given material, if all other factors remain the same one would expect the magnetostriction to depend directly on the magnetic saturation. The Cunico specimen has a low value of magnetostriction (7.5×10^{-6}), a low indentation hardness (R.C. 20) and a high coercive force (445 oersteds) and it therefore contributes strongly to the idea that the classical strain

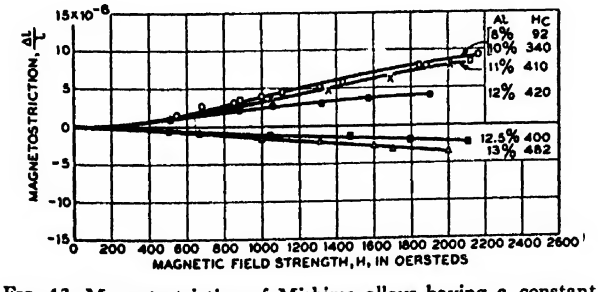


FIG. 13. Magnetostriction of Mishima alloys having a constant nickel content of 29 percent.

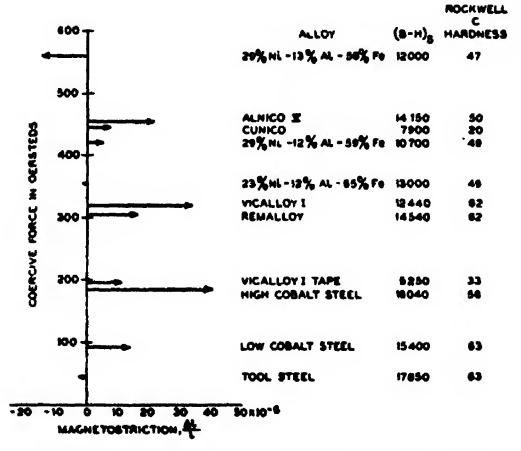


FIG. 14. Magnetostriction of alloys having coercive forces in the range of 50 to 600 oersteds.

[13] V. E. Legg, Bell Sys. Tech. J. 18, 248 (1939).
[14] E. A. Nesbitt, Trans. Am. Inst. Min. Met. Engrs. 166 (1946).

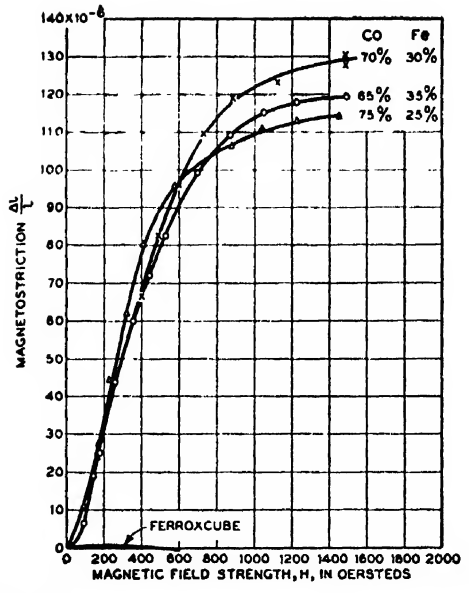


FIG. 15. Magnetostriction of alloys of special interest.

theory does not apply to this alloy. The Vicalloy tape specimen (impregnated in Bakelite) also has low magnetostriction (11×10^{-6}) and low hardness (R.C. 33). The Mishima composition of 23 percent nickel, 12 percent aluminum, and 65 percent iron, cooled from 1300°C at the rate of 5°C per second had the lowest magnetostriction of the alloys actually measured, a value of -1.0×10^{-6} with an accompanying coercive force of 356 oersteds. The magnetostriction of the other Mishima composition (29 percent nickel, 13 percent aluminum and 58 percent iron) is plotted for the sand-cast state, for which the saturation magnetostriction is -15×10^{-6} and the coercive force 560 oersteds.

In Fig. 14 one observes that for the carbon hardening alloys measured there is an approximately linear relation between magnetostriction and coercive force but for the carbon-free alloys measured there is no such relation.

The fourth and final phase of this work was the measurement of several alloys of special interest. Figure 15 shows a magnetostriction curve on the Ferroxcube specimen. The magnetostriction was less than 1×10^{-6} for all values of the field up to 600 oersteds. This result is in accordance with the theory of ferrites in which oxides are present in the proper proportion to obtain approximately zero magnetostriction.[15] Figure 15 also shows curves of magnetostriction obtained on tape specimens (0.002 in.×0.250 in.) of iron-cobalt alloys having 65, 70, and 75 percent cobalt. The tape was given 95 percent cold reduction, wound into a specimen form (2.0 in. I.D., 2.5 in. O.D.) and the turns consolidated by impregnating in Bakelite. Values of 120, 130, and 114×10^{-6} were obtained respectively. A value of 80×10^{-6} had previously been reported[16] on the alloy containing 65 percent cobalt. In our specimens, we have oriented the domains transversely by drastic cold work so that the maximum change in length is observed when the domains are oriented by the applied field. The value of 130×10^{-6} obtained on the 70 percent cobalt alloy is one of the highest values of magnetostriction reported in the literature.

CONCLUSIONS

The curves obtained on Alnico V show that its magnetostriction depends upon heat treatment, the values ranging from 22 to 35×10^{-6}. The highest value of 35×10^{-6} was for the cooling rate which gives optimum permanent magnet properties. If we include the magnetic field treatment, the magnetostrictive values of Alnico V vary from 2.5 to 43×10^{-6} or more than a tenfold variation. The low longitudinal value is due to the fact that most of the domains are oriented either parallel or antiparallel to the magnetization and a 180° reversal of the domains does not contribute to the change in length. The high transverse value is due to the fact that most of the domains are oriented at right angles to the magnetization and each domain contributes to a change in length. The behavior of magnetostrictive specimens of Alnico V which were treated in longitudinal and transverse fields was in accordance with domain theory, as has been pointed out by Hoselitz and McCaig.[17] They concluded from their measurements that the domain magnetization is along the easy crystallographic direction which makes the smallest angle with the axis of anisotropy. However, our data indicate that the domain magnetization is more closely aligned with the direction of the field applied during heat treatment (see Figs. 9 and 10).

The alloys in the immediate vicinity of Alnico V have yielded substantial but not quite as high values of magnetostriction as Alnico V. The alloy containing 6 percent aluminum had a maximum magnetostriction of 25.5×10^{-6}, the standard 8 percent aluminum alloy had a maximum magnetostriction of 35×10^{-6} while the 10 percent aluminum alloy had a maximum value of 19×10^{-6}.

As a result of the investigation of the magnetostriction of permanent magnets having various coercive forces it appears that for alloys hardened by carbon (tool steel, cobalt steels) there is a rough linear relationship between magnetostriction and coercive force for the alloys tested. When we include the carbon-free magnets in the comparison such as Alnico, Vicalloy, Remalloy, Cunico, and the iron-nickel-aluminum system, the linear relation no longer exists. For example Alnico V has over twice the coercive force of Honda steel but its magnetostriction is only 60 percent of that of Honda steel. Likewise the Remalloy and Vicalloy specimens measured had ap-

[15] J. L. Snoek, *New Developments in Magnetic Materials*, (Elsevier Publishing Company, Inc., Amsterdam, 1947).

[16] Y. Masiyama, Sci. Repts. Tohôku Imper. Univ. 21, 394–410 (1932).

[17] Hoselitz and McCaig, Proc. Phys. Soc. 62B, 163 (1949).

proximately the same coercive force but the magnetostriction of the Vicalloy was twice that of Remalloy. Cunico had a low magnetostriction and a low indentation hardness compared with its high coercive force of 445 oersteds. Finally many of the iron-nickel-aluminum alloys had a low magnetostriction combined with a high coercive force and for some of these alloys the magnetostriction actually became zero (Fig. 13).

On the basis of this experimental work one can say that the coercive force of some of the iron-nickel-aluminum alloys does not depend on the magnetostriction. Also, it may be concluded that in the other carbon-free alloys investigated, the magnetostriction is not the only factor operating in these systems to produce the high coercive force, especially in the case of Alnico V since this alloy is directly related to the iron-nickel-aluminum system. For the three carbon-hardening alloys investigated it appears that there may be a close relation between magnetostriction and coercive force.

Recent theories of coercive force have emphasized inhomogeneity of composition and fine particles as well as internal stress. The results obtained in this paper suggest that inhomogeneity of composition or the existence of fine particles is the controlling factor in the iron-nickel-aluminum system and that they may be a factor in the other carbon-free alloys investigated.

ACKNOWLEDGMENTS

The author is greatly indebted to Dr. R. M. Bozorth for discussions and suggestions given during the work, to Miss M. Goertz who assisted in making the measurements, and to Mr. D. H. Wenny who supervised the making of the alloys used.

Improved Ultra-Thin Sectioning of Tissue for Electron Microscopy*·†

JAMES HILLIER
RCA Laboratories, Princeton, New Jersey and Sloan-Kettering Institute for Cancer Research, New York, New York

AND

MARK E. GETTNER
Sloan-Kettering Institute for Cancer Research, New York, New York

(Received March 30, 1950)

The use of a reservoir of liquid to collect ribbons of thin sections as they are cut enables serial sections to be obtained and provides means for detecting variations in thickness. Reducing the effects of static friction and of external vibrations makes it possible to cut long series of sections 0.2μ thick and uniform to better than five percent. Some observations which may throw some light on the mechanics of the cutting process are reported. Mounting the sections without any treatment is shown to be superior to any technique in which the embedding material is removed or replaced. It is shown that the technique provides sections in which a structural resolution of 200A is possible. The remaining artifacts introduced by the sectioning and mounting are easily recognized deformations which do not influence the interpretation.

I. INTRODUCTION

ANY analysis of the requirements for the widespread application of the electron microscope in the fields of biology and medicine leads directly to the necessity of sectioning tissue routinely at thicknesses of the order of 0.2μ, preferably serially. Its acceptance by the cytologists, who will ultimately be the users, will depend entirely on the ability of the technique to provide resolution of structure and freedom from artifacts considerably beyond those obtained by the conventional techniques of histology and cytology. Pease and Baker[1] made a major contribution to this work when they demonstrated that it is feasible to cut sufficiently thin sections by relatively simple modifications of conventional microtomes and techniques. Viewed in retrospect, their subsequent work and that of the workers who followed their technique shows rather large scale artifacts. A simple check with a light microscope shows that these artifacts are of considerably greater magnitude than would be expected from the fixation and imbedding used. It must be assumed therefore, that the artifacts were introduced by the sectioning technique or the subsequent manipulation of the sections. The research reported in this paper involves a careful analysis of the Pease and Baker technique with the result that many of the sources of artifacts have been identified and removed.

The complete problem of examining tissues with the electron microscope involves: (1) the selection and fixation of the tissue, (2) embedding, (3) sectioning, and (4) mounting of sections. The first two steps include a number of variables about which no information can be

* This work was sponsored in part jointly by the ONR under Contract No. N6-ori-99 Task Order I and the AEC and in part by the Lillian Babbitt Hyde Foundation. The electron microscope used was supplied through the kindness of New York University.

† Based on a paper presented at the Washington D. C. meeting of the Electron Microscope Society of America, October, 1949 [J. App. Phys. 21, 67A (1950)] supplemented by more recent work.

[1] D. C. Pease and R. F. Baker, Proc. Soc. Exper. Biol. and Med. 67, 470–474 (1948).

obtained until the last two are developed to the point where routinely reproducible results are possible. Thus the work described in the following was all carried out on a set of specimen blocks obtained from osmium-perfused mouse liver. Actually the livers from only two mice were used and all the blocks from each liver were embedded simultaneously by the technique described in Pease and Baker's original work.[1]

II. SECTIONING: BLOCK ADVANCE

In spite of its multiplicity of moving parts, the Spencer rotary microtome (No. 820) has proved to be ideal for the cutting of thin sections. In this machine the knife is stationary and the block is moved vertically to make the cuts and advanced horizontally between cuts. In the mechanical design of the Spencer microtome, these two motions are intimately related and will be considered in the discussion concerning the advancing of the block. It was found that for very thin sections which require little cutting force the actual sectioning (that is, the action of the knife on the block) is not related to the advancing of the block except under some rather special conditions.

The reservoir for collecting and spreading the specimens serially, described by Gettner and Hillier[2,3] also proved very useful for testing the operation of the advance mechanism. It was found that for a given block, knife, cutting speed, and reservoir liquid the length of the floating section (in the direction of cutting) is a direct function of the thickness of the cut. The shortening of the sectioning is first appreciable at 0.3μ and increases rapidly with decreasing thickness, usually being greater than 50 percent for sections of 0.1μ thickness. Thus, the lengths of successive sections observed at a magnification of 27, as they appear, provides a sensitive criterion of the uniformity of thickness. By this means it has been possible to identify several sources of variations in the advance mechanism: (1) static friction, (2) vibrations of external origin, and (3) thermal expansion. In an unchanged Spencer microtome, operated by hand, these effects produce variations in thickness of the order of 0.2μ which while insignificant in normal sectioning are much too great for ultra-thin sectioning. *It is to be noted, from the discussion which follows, that modification of the method or magnitude of advance has no effect on these sources of variations (except possibly a slightly detrimental one).* Further refinements are necessary therefore, if sections of uniform thickness at the 0.1μ level are to be cut properly.

FIG. 4. Normal mouse liver cut at 0.2μ. Embedding was completely replaced with collodion. Micrographed with a double objective with limiting aperture set at 4×10^{-3} radian. This section demonstrates a uniform compression of approximately 40 percent which means that the actual specimen is more than 0.3μ thick. Destructive "knife marks" due to edge defects 8 and 2μ long are visible. All of the micrographs presented in this paper were with a standard RCA Model u.c. operating at 50 kv. Magnification approximately 900×.

The advance mechanism of the Spencer microtome consists of a series of mechanical reductions.[4] At the top of each cutting cycle, a pawl actuated by the drive wheel engages a toothed ratchet wheel and advances it one tooth for the unit advance. Coaxial with, and rigidly attached to the wheel is an accurately threaded rod which engages a threaded nut. The nut in turn is rigidly attached to a pointer which rides against an inclined plane, an integral part of the specimen block holding and moving system. The forward motion of this last unit is guided by a set of horizontal slides which also support it. A spring keeps the inclined plane in contact with the pointer. The cutting is accomplished by moving the block and its support, including the spring, inclined plane, and horizontal slides, in a set of vertical slides. The cutting motion is sinusoidal and has an amplitude of 5 cm. Since the pointer is stationary, it traces a line on the inclined plane approximately vertical and approximately perpendicular to the inclination. If the normal to the inclined plane were in the horizontal plane, or, more precisely, perpendicular to the direction defined by the vertical slides, the true motion of the specimen block would be parallel to the vertical slides. Under this condition there would be no motion of the block support relative to the horizontal slides during the cutting movement. Actually, it is an important feature of the Spencer microtome that the normal to the inclined plane usually makes a small angle to the horizontal. This means that, as the block makes its vertical excursion of 5 cm, it also makes a horizontal excursion of several microns (300 in our case). Thus, there is relative motion in the horizontal slides some time before and during the actual cutting. By this simple means, the large varia-

[2] M. E. Gettner and J. Hillier, Science (submitted for publication).

[3] M. E. Gettner and J. Hillier, J. App. Phys. 21, 68A (1950).

[4] O. W. Richards, *The Efficient Use and Proper Care of the Microtome* (American Optical Company, Buffalo, New York, 1949), second edition, p. 84.

tions in the advance of the block caused by static friction in the horizontal ways are almost completely eliminated. Some crude tests indicated that the force required to overcome the static friction in the horizontal slides was such as to produce a deformation of between 1 and 2μ in the inclined plane and block support. Thus if the horizontal excursion were eliminated and the advance set at 0.1μ, the specimen would still advance in 1 to 2μ steps and be cut only after every 10 to 20 cutting cycles. The horizontal excursion also means that any "play" in the vertical slides or any flexibility in any other associated parts is taken up by the same amount in each cycle.

It was found that for ultra-thin sections static friction in the horizontal slides still produced greater variations than were desirable. This was due to the fact that the relative motion in the horizontal slide stopped and reversed at the top and bottom of each cutting cycle allowing static friction to introduce variations which were not entirely removed by the time the block reached the knife. The addition of considerable extra weight in the modification of the inclined plane aggravated this situation. The variations in section thickness were then of the order of 0.05 to 0.1μ. In view of the reduced total advance in the modified system, it was possible to eliminate these variations by removing about 90 percent of the weight of the block support and modified inclined plane from the horizontal slides. To do this, a short coil spring of the proper tension was inserted between the rear point of attachment of the main spring and the lower brace for the inclined plane. While this was satisfactory, it was found more convenient to use a leaf spring mounted on the top of a new block which also serves to secure the back end of the main spring. The lifting force is applied to the block support by a bent rod which hooks under the lower brace of the inclined plane. The upper end of this rod is threaded and secured to the end of the leaf spring by a small nut. Thus the lifting force can be adjusted. The horizontal excursion of the block and the total advance are permitted by the fact that the leaf spring can twist and that neither end of the bent rod is rigidly attached.

The modification of the inclined plane used follows the principles described by Pease and Baker. There are, however, a number of specific points of difference. The modification consists of machined iron block which slips over the original inclined plane. It is held in place by eight pointed set screws which grip the original plane behind its top and bottom edges. There is close contact between the top and bottom edges of the original plane and the block but none in the central region. The new inclined plane is a selected piece of plate glass which is flat to within two fringes of sodium light. It is fastened to the iron block by embedding in Woods metal. Before embedding, the glass plate was adjusted to the proper position, by what amounts to three leveling screws threaded into the block.

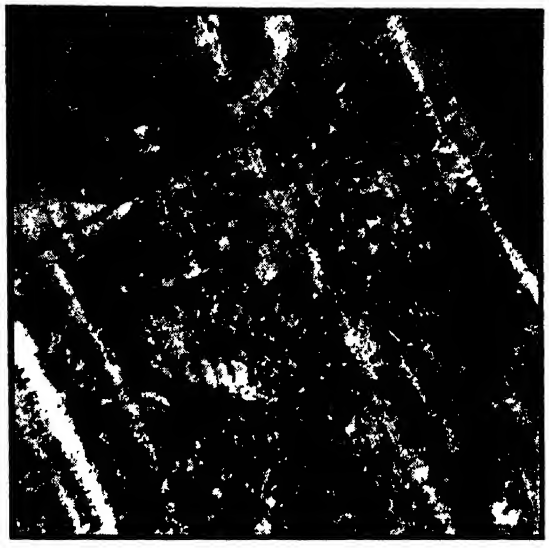


FIG. 2. Normal mouse liver cut at 0.2μ. Embedding was left intact. Micrographed with conventional objective but with focal length of approximately 8 mm and limiting aperture set at 3×10^{-3} radian. This shows a number of fine folds perpendicular to the direction of cutting caused, in this case, by the greater compression which occurred in the thinner strip which separates the two rows of folds. Note also that the "knife marks" are more evident than in Fig. 1. Magnification approximately 900×.

The unit of advance was set at 0.02μ. This low value was not chosen with any thought that sections of this thickness could be cut, but rather to overcome possible variations due to static friction in the slides that support and guide the pointer between the nut and the thread and in the end bearings of the thread and drive wheel. With this arrangement, sections of thickness 0.2μ which have been found most suitable require a rotation of the drive wheel and thread of 16.4° and a movement of the pointer of 20μ. Both of these movements are large enough to be accurately reproducible. This means also reduces the effects of inaccuracies in the threaded rod.

Variations of section thickness equal in magnitude to those put in by the effects of static friction are introduced by vibration of the microtome or of its support. The main trouble in this respect was traced to the irregularities of manual operation of the instrument. It was successfully overcome by adding a motor drive. A small 1/15 hp motor was used with a built-in reduction gear that gave 30 r.p.m. The motor was ballasted and shock mounted on sponge rubber separately from the microtome. In order to avoid transmission of vibrations through the coupling to the microtome and also to avoid difficulties of accurately aligning the drive shafts of the microtome and the motor, a special coupling was devised. The power is transmitted through a strip of Neoprene (approx. $\frac{1}{16}\times\frac{3}{8}\times1\frac{1}{2}$ in.) from the end of a 3-in. long bar clamped to the motor output shaft to a pin set in the microtome drive wheel, also 3 in. from the axis. The Neoprene seems to have the right degree of elasticity with the proper damping for the arrangement

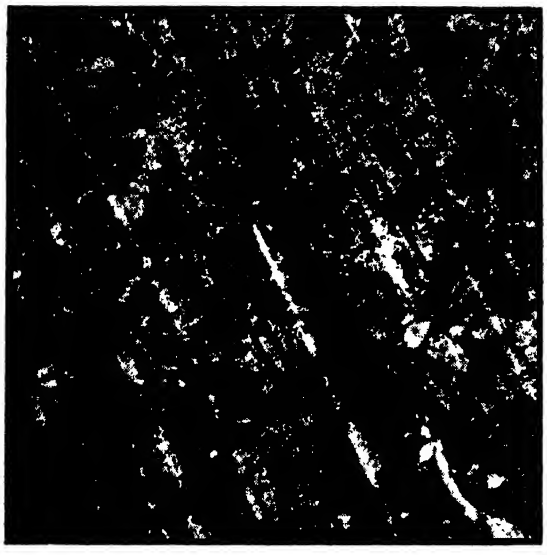


FIG. 3. All conditions identical with those of Fig. 2. Here a dense granule has been carried by the knife splitting the section until it was left on the section. Note that some of the denser "knife marks" appear to consist of a homogeneous material (paraffin?) which has been left on the specimen. Magnification approximately 900×.

used. Direct coupling by means of a wire loop or coupling through a coil spring under tension produced periodic variations in thickness.

A brass plate was added to the drive wheel of the microtome to counterbalance accurately the change in weight of the horizontal member caused by the addition of the modifications.

The instrument remains extremely sensitive to externally produced vibrations. In fact, resting a finger very lightly on any part of it will introduce variations in thickness amounting to 0.05 to 0.1μ. Building vibrations caused by nearby traffic also cause variations. Although it was not done in this work, it would be desirable to mount the instrument on a high inertia vibration-free support.

Thermal changes in the instrument are the only remaining major source of thickness variations. The time constant of these changes is so long, however, that with motor drive, the effect is more one of a disturbance of calibration. Thus, it cannot be detected directly by observation of the uniformity of compression as the sections appear on the liquid surface. It can be detected, however, by setting the advance to zero and observing the cutting through a low power microscope. A continuation of cutting or delay in returning to cutting when the advance is set to a finite value indicates the presence of thermal changes. More recently it has been found that an accurate estimate of the direction and amount of thermal drift can be obtained by the following procedure. The machine is allowed to cut a series of sections of desired thickness, say 0.2μ, then within one cycle the manual advance crank is turned backward one turn while the machine is allowed to continue operating. This represents a retraction of the block by 25 0.2μ sections. The knife is observed and the cutting cycles counted until the first new section is cut. This should happen on the 26th cycle. Any deviation will indicate the presence of thermal drifts. Furthermore, the amount and direction of the deviation will enable the true thickness of the cut to be calculated to within four percent in this example. More than one turn of the crank can be used if higher accuracy is desired. The method has the advantage that any disturbance introduced by touching the instrument is completely damped out by the time a new cut is made.

The most important source of thermal changes has been from the use of a focused light to illuminate the edge during observation. This practice has now been discontinued in favor of using a large fluorescent light placed some distance above the instrument. Ordinary changes in room temperature introduce little effect though it is essential that the instrument be placed in a location free from strong air currents.

A possible source of variations which has not yet been studied is in the thickness of the oil film which lubricates the vertical ways. Variations in the thickness of this film of 40A would be detectable under optimum conditions. Obviously such changes could occur. Furthermore, the "working in" of an oil film of different thickness could produce periodic variations. In practice, very little lubrication (Pike oil) has been used and the microtome is run at zero setting for several minutes before sections are cut. These are mainly precautionary measures and may prove unnecessary.

A final point which should be mentioned in connection with the function of the advance mechanism is in regard to the adjustment of the slides. This proved to be simpler than was expected because the cutting force required for the thin sections and very small blocks is very small. This force is never sufficient to lift the weight of the block supporting mechanism from the lower of the horizontal slides, nor to separate the rear of the vertical slides against the tension of the main spring. Thus, since a transference of the reactive force from one slide to the other never occurs in either set, there is no need to adjust the slides so that there would be no displacement of the block if such a transfer were to occur. In this work, the slides are adjusted so as to be very free running and are somewhat looser than would be considered good practice in ordinary microtomy.

III. SECTIONING: CUTTING

The interaction of the knife with the specimen which constitutes the cutting of sections is, as yet, a very poorly understood phenomenon. From this work it has become evident that defects in a knife-edge produce observable effects if they have dimensions of the same order as the thickness of the section cut. Thus it would appear that the cutting of ultra-thin sections will provide information regarding the nature of the edge at a level at least an order of magnitude below what was

previously possible and thus provide a very profitable approach to this fundamental problem.

It is apparent that the quality of the sections will be a complicated function of many variables, including in particular, the nature of the tissue, the fixation, the dehydration, the embedding, the sharpness and cleanliness of the edge, the bevel and clearance angles, the speed of cutting, the temperature, and the properties of the liquid in the reservoir. An attempt was made to reduce the number of variables by using only one specimen of fixed tissue cut into a number of pieces which were dehydrated and embedded simultaneously. Unfortunately, this leaves an uncomfortably large number of variables, at least two of which cannot be determined independently—namely, the sharpness and cleanliness of the edge. Nevertheless, some conclusions can be drawn from the observations already made.

For the work shown here, commercially sharpened knives were used throughout. It is recognized that such edges represent a compromise between quality and cost. More recently, hand-sharpened edges have been providing much superior results. The edges of the blades are examined at 400× by the scattered light technique described by Richards[2] and the facets by reflected light.

The fundamental defects in the sections obtained in this work consists of a uniform microscopic compression (Fig. 1), an accordian-like folding parallel to the cutting edge (Fig. 2) strips of partially destroyed tissue which run parallel to the direction of cutting (Fig. 1), and local distortions due to different cutting properties of different areas of the tissue. The last defect can vary in magnitude from the picking up by the blade of small hard granules which cut the sections until they are dropped (Fig. 3) to the folding of large areas. In passing, it should be mentioned that in material which is much harder than the embedding, there is a characteristic "chattering" phenomenon involving a rather uniform rearrangement of the tissue in 5 to 15μ steps.

The compression which appears as a microscopically uniform contraction in the direction of cutting is dependent on the nature of the tissue and the thickness of the section. It is definitely not dependent on the rake angle over the 6° range used in the tests. This is contrary to the experience of Baker and Pease.[5] This may be due to the presence of the liquid surface.

The accordion-like folding is peculiar to thin sections and does not appear in sections thicker than 0.3μ. It appears that the great flexibility of the sections permit them to fold whenever they encounter mechanical resistance beyond the cutting edge of the knife. As would be expected, the persistence of the folds is very sensitive to the cleanliness of the knife and to the surface tension and level of the liquid. It has been found necessary to adjust the surface tension and level of the liquid for many of the blocks. Observation of the sections at approximately 30× as they are cut provides a good criterion for these adjustments. It has been found that good sections appear suddenly and fully extended on the surface of the liquid. Moreover, they are transparent and nearly invisible. Poor sections appear to slide slowly on to the liquid from a folded condition near the edge of the blade and appear somewhat translucent.

A very special type of folding which cannot be removed is caused by a defect in the knife which cuts a thin track through the specimen without causing any destruction. The thinner line is more compressed with the result that the adjacent areas must fold to compensate for the change in length of the section (Fig. 2).

Large defects in the knife will simply split the section, but more often, a characteristic line of destruction is produced (Fig. 1). The latter occurs for nicks or other defects which are isolated and occupy more than 1μ of the length of the edge. Defects occupying a length of the edge approximately twice the thickness of the section appear to produce slight local deformations of the structure (Fig. 4).

Observations of long series of sections cut at one point of the knife indicate that the dulling is a step-by-step process involving an increase in the number of larger defects in the knife. A relationship has also been observed between the sharpness of the blade and the frequency with which it must be cleaned. A slow decrease in the lengths of successive sections and an immediate return to the original length upon cleaning is indicative of an edge that is in need of sharpening. Examination of sections produced by such an edge show a greatly in-

FIG. 4. A portion of the same section as shown in Fig. 1 taken at higher electronic magnification and with the same instrumental conditions. The nuclear membrane appears to be slightly deformed by a defect in the knife roughly estimated to be 0.5μ long. In comparing this with Fig. 6, note the thicknesses of the cell wall, the nuclear membrane, and the intracellular fibers. Note also that the granules in the cytoplasm appear hollow with dense walls. Magnification approximately 3,700×.

[5] R. F. Baker and D. C. Pease, J. App. Phys. 20, 480 (1949).

creased number of lines of destruction which are also somewhat wider than obtained with a sharp edge.

Another interesting observation on the nature of the edge can be made by retracting the block a few section thicknesses and moving the blade, parallel to itself, to a new position. If the first cuts are observed, it is found that they are invariably narrow strips, all parallel to the direction of cutting, which grow wider on each successive cut until a complete section is obtained. This usually requires a total advance of the block of something more than a micron. This clearly indicates that the cutting edge is not a straight line but is corrugated with an amplitude of approximately 1μ.

IV. MOUNTING AND MICROGRAPHY OF SECTIONS

The handling of the specimens after they have been cut depends to a certain extent on their thickness, the nature of the tissue, and the problem under investigation. From our work, it appears that earlier workers have grossly underestimated the delicacy of the cut sections and that it is in this stage of the work that most of their artifacts have been introduced.

A technique described by Pease and Baker[3] involved mounting the section on a glass slide, after it had been spread on dioxane, and allowing it to dry. It was then washed with benzene to dissolve out the paraffin. This was followed with 0.2 percent collodion in amyl acetate which washed out the collodion embedding but left in a thin layer to support the specimen. The section was then floated off the slide on to clean water and mounted on a 200-mesh screen for examination in the microscope.

This technique was found to introduce artifacts in two ways. While the dioxane is an excellent means of spreading and flattening the section at the concentrations used, it nevertheless introduces considerable distortions in a way that is not completely understood. The use of the dilute collodion solution meant that the final film was much thinner than the section and thus, in the final stages of drying, surface forces are allowed to act on the very fragile structures of the section. This produced gross distortion of all but the coarsest structures (Fig. 5). The technique was then modified to omit the dioxane and to use a two percent solution of collodion leaving enough of the solution on the section to dry down to a film that was slightly thicker than the section. This eliminated the distortions due to surface forces, but greatly reduced the contrast obtainable and introduced difficulties in making the sections adhere to the specimen screens.

FIG. 5. Section cut at 0.2μ from which the embedding was entirely removed and replaced with collodion in amyl acetate of such concentration and amount as to produce a dried film of approximately 400A thickness. That there are large scale distortions is obvious from comparison with 2, 3, and 6. No objective aperture. Magnification approximately 900×.

FIG. 6. A field selected for comparison with Fig. 4. The conditions of sectioning and micrography are the same with the exception that the embedding was not removed. Magnification approximately 3000×.

The contrast problem was overcome by raising the specimen to give an objective focal length of approximately 7 mm and by using a standard (50μ diam.) objective aperture resulting in an angular aperture of 3.5×10^{-3} radian.

This modified technique was finally abandoned when serial sections showed that merely dissolving the embedding material at any time after sectioning leaves many of the finer structures free to move. (Compare Figs. 4 and 6.)

The mounting technique now used involves no treatment of the section other than a mechanical mounting on a specimen screen (Figs. 6 and 7). A piece of 200-mesh copper screen (Elektromesh) 2×3 in. is etched in dilute nitric acid until it has at least 50 percent open area. A 300A thick collodion membrane is then mounted on this screen in the conventional way and dried. The sections

are mounted on suitable pieces of this screen by lifting them directly from the collecting reservoir. It is important that the prepared screen be lowered in to the liquid in a nearly vertical position but with the collodion membrane on the underside and that it be moved slowly through the liquid with film-side facing the direction of motion. After the sections are lifted from the liquid, the blank side of the screen is immediately blotted with hard dry filter paper and as much liquid as possible is drawn out from between the sections and the film.

The use of the extra film has successfully overcome the difficulty of anchoring the sections to the screen. In these sections there remains a problem of shrinkage under moderate electron bombardment. It can be avoided by raising the intensity quite slowly on each new field. Shrinkage of the section from around the screen openings can be avoided by placing the mount in the microscope with the sections facing the electron source, so that the whole area bombarded is seasoned uniformly.

The use of some dioxane (less than 50 percent) in the reservoir has been found to improve the flattening of the section in some cases, but requires that they be treated much more carefully under bombardment.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Using the techniques described above, a large number of micrographs of the osmium-perfused liver have been obtained in which the structures have been consistent. The structure resolution is between 200 and 300A for the 0.2μ sections used. It is obvious that many new structures are being observed though their significance and even their reality are yet to be determined. Several artifacts introduced by the sectioning technique are still present. These include compression, folding, knife marks, shrinkage under bombardment, and occasional removal or shifting of material. However, all of these defects are quite obvious and do not interfere with the correct reading of the micrographs. It would be desirable, of course, to eliminate these as they detract from the esthetic value of the micrographs and can be annoying.

As indicated in the beginning, this investigation concerned only the factors involved in the cutting of uniformly thick sections and in the mounting for electron microscopic examination. The tissue, fixation, and

FIG. 7. A representative field of a 0.2μ section mounted with the embedding intact. Magnification approximately 1800×.

embedding was maintained as constant as is possible in biological preparations. A type of tissue and a technique which had already been shown to be quite favorable for the problem were chosen. It is to be noted, however, that the refinements introduced are only slightly dependent on the nature of the preparation. This is not true of the cutting properties which can vary widely according to the nature of the tissue of the fixative, or of the embedding.

A number of other preparations have now been sectioned with varying degrees of success. However, while experience in this respect has not suggested any modifications in the techniques described above, it has emphasized the necessity of adjusting the embedding very carefully to the properties of the specimen and the technique of fixation and the need for sharper knives.

ACKNOWLEDGMENTS

The authors wish to thank Drs. J. J. Biesele and J. A. Jacquez for providing carefully prepared specimens and for their valuable discussions of the resulting micrographs. Thanks are also due Miss G. Cohn for her assistance in the embedding and mounting of the blocks.

Reflection and Transmission of Electromagnetic Waves by Thin Curved Shells*

JOSEPH B. KELLER
Institute for Mathematics and Mechanics, New York University, New York, New York
(Received March 31, 1950)

The scattering of an arbitrary electromagnetic field by a conducting or non-conducting obstacle is investigated. The differential equations and boundary conditions satisfied by the field are transformed into a pair of inhomogeneous linear integro-differential equations for E and H. For an obstacle which is a thin shell of constant thickness h, a formal procedure for obtaining a solution of these equations as power series in h is given. The lowest order term in this solution is the incident field. An explicit expression for the next term is found in the form of a surface integral. This integral is evaluated approximately by the method of stationary phase. The physical properties of the solution are examined in detail, and satisfactory agreement is found with many results previously obtained by other methods.

1. INTRODUCTION

WE wish to find the electromagnetic field which results when a given incident field strikes an obstacle—more precisely the field produced by a given source in the presence of an obstacle. For some special obstacles the problem has been solved exactly, and for arbitrary obstacles several approximate methods are available (geometrical optics,[1-3] Kirchhoff method,[3,4,5-7] Rayleigh scattering formula). Recently, Professor H. Primakoff and the author treated the corresponding acoustic problem for thin shell obstacles by a method which takes account of propagation within the obstacle and includes the results of the methods listed above.[3] In this paper the same method is applied to electromagnetic scattering by thin shells.

In Section 2 the problem is formulated and converted into an inhomogeneous integral equation. In Section 3 an approximate solution is obtained for an obstacle which is a thin shell. In Section 4 the integral representing the approximate solution is evaluated and in Section 5 the solution is examined and compared with other results. Many of the omitted details may be found elsewhere.[3,8]

2. DERIVATION OF THE INTEGRAL EQUATION

An obstacle (b) is immersed in an infinite medium (a), both being homogeneous and isotropic. A periodic source ($e^{-i\omega t}$) is located in (a). Outside the source the electric field $E(r)$ satisfies the reduced wave equation

$$(\nabla^2+k^2)E(r)=0. \tag{1}$$

Here $k^2=(\omega^2/c^2)\epsilon'\mu$; $\epsilon'=\epsilon+(4\pi i\sigma/\omega)$; $k\equiv k_a$ in (a); $k\equiv k_b$ in (b). (We shall write equations for E only; the corresponding equations for the magnetic field $H(r)$ are obtained by interchanging E with H and ϵ' with $-\mu$.) At the boundary between the two media the field satisfies the discontinuity conditions

$$E_t{}^a=E_t{}^b; \quad \epsilon_a'E_n{}^a=\epsilon_b'E_n{}^b. \tag{2}$$

Here n and t indicate normal and tangential components, and a, b indicate sides of the boundary, respectively. At infinity the scattered field $E-E_0$ satisfies a radiation condition. The source is characterized by $E_0(r)$, the field which would be produced in the absence of the obstacle. (We assume the obstacle does not affect the source.)

All these conditions and equations will be transformed into an integro-differential equation with the aid of a Green's function $G(r,r')$ which satisfies Eq. (1), with $k=k_a$ everywhere, with respect to the variable r'. At $r'=r$, G has a singularity of the form $1/(|r-r'|)$, and at infinity it satisfies a radiation condition. We apply Green's theorem to $G(r,r')$ and $E(r')-E_0(r')$, assuming $E-E_0$ has no singularities. We first integrate over the obstacle V bounded by S', then over the rest of space, and add, obtaining (r' is the integration variable)

$$E(r)-E_0(r)=\frac{1}{4\pi}\int_{S'}\left\{(E^a-E^b)\frac{\partial G}{\partial n}-G\left(\frac{\partial E^a}{\partial n}-\frac{\partial E^b}{\partial n}\right)\right\}ds+\frac{k_b{}^2-k_a{}^2}{4\pi}\int_V EG\,dv. \tag{3}$$

* The research reported in this document has been made possible through support and sponsorship extended in part by the Geophysical Research Directorate of the Air Force Cambridge Research Laboratories, AMC, Contracts W28-099-ac-172 and with Washington Square College, New York University.

[1] H. J. Riblet and C. B. Barker, J. App. Phys. 19, 63 (1948).

[2] J. B. Keller and H. B. Keller, J. Opt. Soc. Am. 40, 48 (1950).

[3] H. Primakoff and J. B. Keller, J. Acous. Soc. Am. 19, 820 (1947).

[4] J. B. Keller, "Reflection of electromagnetic waves," Special Report No. 172-1 to the Watson Laboratories; New York University, Washington Square College Mathematics Research Group, (December, 1946).

[5] Lord Rayleigh, *Theory of Sound, II* (Dover Publications, New York, 1945), pp. 118-126.

[6] Willis, "Reflection and refraction of ASDIC beams from various bodies and surfaces," British Admiralty Report (1943). (Integrals evaluated by Fresnel zone method.)

[7] "The scattering of radiation from rectangular planes, half-cylinders, hemispheres, and airplanes," report submitted to the Evans Signal Laboratory by the Moore School of Electrical Engineering, University of Pennsylvania (October, 1943). [Only linear terms retained in the phase (Frauenhofer approximation) which yielded easy integrals but some unreasonable results.]

[8] J. B. Keller, "Reflection and transmission of electromagnetic waves by thin curved shells," thesis, New York University (April, 1948).

To evaluate the first difference occurring in the integrand in Eq. (3), we use Eq. (2). We find (n is the unit normal)

$$E^a-E^b=[(\epsilon_b'/\epsilon_a')-1]E_n{}^b n. \tag{4}$$

The evaluation of the second difference is more difficult and requires the use of Eq. (2), Maxwell's equations and the use of curvilinear coordinates. The lines of curvature on the boundary surface S', designated as $u_1=$const., $u_2=$const. are employed. We finally obtain (see reference 8 for details)

$$\frac{\partial E_n{}^b}{\partial n}-\frac{\partial E_n{}^a}{\partial n}=\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)E_n{}^b\cdot 2G_m, \tag{5}$$

$$\frac{\partial E_1{}^b}{\partial n}-\frac{\partial E_1{}^a}{\partial n}=\frac{i\omega}{c}(\mu_b-\mu_a)H_2{}^b+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\frac{\partial E_n{}^b}{\partial u_1}, \tag{6}$$

$$\frac{\partial E_2{}^b}{\partial n}-\frac{\partial E_2{}^a}{\partial n}=-\frac{i\omega}{c}(\mu_b-\mu_a)H_1{}^b+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\frac{\partial E_n{}^b}{\partial u_2}. \tag{7}$$

Here u_1 and u_2 represent arc length along the lines of curvature; E_1 and E_2 represent components along these lines and G_m is the mean curvature of the surface S'. (A curvature is positive if the center of curvature and positive normal are on opposite sides of S'.)

We now use Eqs. (4)–(7) in Eq. (3) and obtain (i_1, i_2 are unit vectors along the lines of curvature)

$$E(r)=E_0(r)+\frac{1}{4\pi}\int_{S'}\left\{\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)E_n{}^b\left(\frac{\partial G}{\partial n}+2GG_m\right)n+G\left[\frac{i\omega}{c}(\mu_b-\mu_a)(i_1H_2{}^b-i_2H_1{}^b)\right.\right.$$
$$\left.\left.+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\left(i_1\frac{\partial E_n{}^b}{\partial u_1}+i_2\frac{\partial E_n{}^b}{\partial u_2}\right)\right]\right\}dS+\frac{k_b{}^2-k_a{}^2}{4\pi}\int_V EG\,dV. \tag{8}$$

Equation (8) is the fundamental integro-differential equation for the field in the presence of an obstacle. The integrals represent the scattered field, and obviously vanish when the properties of the two media are the same.

3. SOLUTION FOR THIN SHELL OBSTACLES

Equation (8) may be simplified and solved if the obstacle is a thin shell of constant thickness h. Then S' consists of two parallel parts S_I and S_{II}, neglecting the edges. If r is exterior to the shell we may expand the integrands of Eq. (8) in Taylor series about points of S_I. Then Eq. (8) becomes approximately, for r in medium (a) (n is a unit normal to S_I pointing into the obstacle)

$$E(r)=E_0(r)+\frac{1}{4\pi}\int_{S_I}\sum_{m=0}^{\infty}\frac{h^{m+1}}{(m+1)!}\frac{\partial^m}{\partial n^m}\left[\frac{\partial}{\partial n}\left\{\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)\left(\frac{\partial G}{\partial n}+2GG_m\right)E_n{}^b n\right.\right.$$
$$\left.\left.+\frac{i\omega}{c}(\mu_b-\mu_a)G(i_1H_2{}^b-i_2H_1{}^b)+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)G\left(i_1\frac{\partial E_n{}^b}{\partial u_1}+i_2\frac{\partial E_n{}^b}{\partial u_2}\right)\right\}+(k_b{}^2-k_a{}^2)GE^b\right]dS. \tag{9}$$

If the edges are taken into account, we obtain another similar line integral over the curve bounding S_I in Eq. (9).

To solve Eq. (9) we represent the field in medium (a) as a power series in h,

$$E(r)=\sum_{\nu=0}^{\infty}h^\nu E_{(\nu)}(r); \tag{10}$$

insert this into Eq. (9) and equate coefficients of like powers of h. It is to be understood that all quantities in Eq. (9) evaluated on side b are to be replaced by appropriate quantities on side a. We then find that $E_{(0)}=E_0$, and that all succeeding terms can be determined successively as surface integrals.

For $E_{(1)}(r)$ we obtain

$$E_{(1)}(r)=\frac{1}{4\pi}\int_{S_I}\left\{n\left[\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)\left(\frac{\partial^2 G}{\partial n^2}+2G_m\frac{\partial G}{\partial n}\right)E_{n(0)}{}^b+\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)\left(\frac{\partial G}{\partial n}+2GG_m\right)\frac{\partial E_{n(0)}{}^b}{\partial n}+(k_b{}^2-k_a{}^2)GE_{n(0)}{}^b\right]\right.$$
$$+i_1\left[\frac{i\omega}{c}(\mu_b-\mu_a)\right)\left(\frac{\partial G}{\partial n}H_{2(0)}{}^b+G\frac{\partial H_{2(0)}{}^b}{\partial n}\right)+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\left(\frac{\partial G}{\partial n}\frac{\partial E_{n(0)}{}^b}{\partial u_1}+G\frac{\partial^2 E_{n(0)}{}^b}{\partial n\partial u_1}\right)+(k_b{}^2-k_a{}^2)GE_{1(0)}{}^b\Big]$$
$$\left.+i_2\left[-\frac{i\omega}{c}(\mu_b-\mu_a)\left(\frac{\partial G}{\partial n}H_{1(0)}{}^b+G\frac{\partial H_{1(0)}{}^b}{\partial n}\right)+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\left(\frac{\partial G}{\partial n}\frac{\partial E_{n(0)}{}^b}{\partial u_2}+G\frac{\partial^2 E_{n(0)}{}^b}{\partial n\partial u_2}\right)+(k_b{}^2-k_a{}^2)GE_{2(0)}{}^b\right]\right\}dS. \tag{11}$$

The integrand involves the field and its derivatives on the b side of S_I, and these must be replaced, using

Eqs. (4)–(7), by corresponding quantities on the a side. Equation (11) then becomes

$$E_{(1)}(r)=\frac{1}{4\pi}\int_{S_I}\Big\{n\Big[E_{n(0)^a}\Big(1-\frac{\epsilon_a'}{\epsilon_b'}\Big)\Big(\frac{\partial^2 G}{\partial n^2}+2G_m\frac{\partial G}{\partial n}\Big)+\Big(\frac{\epsilon_b'}{\epsilon_a'}-1\Big)\Big(\frac{\partial G}{\partial n}+2GG_m\Big)\Big(\frac{\partial E_{n(0)^a}}{\partial n}+2G_mE_{n(0)^a}$$
$$\times\Big[1-\frac{\epsilon_a'}{\epsilon_b'}\Big]\Big)+(k_b^2-k_a^2)\frac{\epsilon_a'}{\epsilon_b'}GE_{n(0)^a}\Big]+i_1\Big[\frac{i\omega}{c}(\mu_b-\mu_a)\Big(\frac{\partial G}{\partial n}H_{2(0)^a}+G\frac{\partial H_{2(0)^a}}{\partial n}$$
$$+\frac{i\omega}{c}G[\epsilon_b'-\epsilon_a']E_{1(0)^a}+G\Big[\frac{\mu_a}{\mu_b}-1\Big]\frac{\partial H_{n(0)^a}}{\partial u_2}+\Big(\frac{\epsilon_a'}{\epsilon_b'}-1\Big)\frac{\partial G}{\partial n}\frac{\partial E_{n(0)^a}}{\partial u_1}+G\Big(1-\frac{\epsilon_b'}{\epsilon_a'}\Big)$$
$$\times\Big(\frac{\partial^2 E_{n(0)^a}}{\partial u_1\partial n}+2G_m\Big[1-\frac{\epsilon_a'}{\epsilon_b'}\Big]\frac{\partial E_{n(0)^a}}{\partial u_1}\Big)+(k_b^2-k_a^2)GE_{1(0)^a}\Big]+i_2\Big[-\frac{i\omega}{c}(\mu_b-\mu_a)\Big(\frac{\partial G}{\partial n}H_{1(0)^a}$$
$$+G\frac{\partial H_{1(0)^a}}{\partial n}-\frac{i\omega}{c}G[\epsilon_b'-\epsilon_a']E_{2(0)^a}+G\Big[\frac{\mu_a}{\mu_b}-1\Big]\frac{\partial H_{n(0)^a}}{\partial u_1}\Big)+\Big(\frac{\epsilon_a'}{\epsilon_b'}-1\Big)\frac{\partial G}{\partial n}\frac{\partial E_{n(0)^a}}{\partial u_2}+G\Big(1-\frac{\epsilon_b'}{\epsilon_a'}\Big)$$
$$\times\Big(\frac{\partial^2 E_{n(0)^a}}{\partial u_2\partial n}+2G_m\Big[1-\frac{\epsilon_a'}{\epsilon_b'}\Big]\frac{\partial E_{n(0)^a}}{\partial u_2}\Big)+(k_b^2-k_a^2)GE_{2(0)^a}\Big]\Big\}dS. \quad (12)$$

Equation (12) represents the scattered field to first order in h in medium (a). Only the integral over the edges has been neglected in obtaining this result. In the same way, the higher order terms in the solution may be obtained.

4. EVALUATION OF $E_{(1)}(r)$

To evaluate the integral for $E_{(1)}(r)$ in Eq. (12) we select for $G(r, r')$ the Green's function for infinite space:

$$G(r,r')=\frac{\exp(ik_a|r-r'|)}{|r-r'|}. \quad (13)$$

We assume that the incident field is given by

$$E_0(r)=\mathcal{E}(r)\exp(ik_a|r-r_0|), \quad (14)$$

where $\mathcal{E}(r)$ is slowly varying compared to the exponential factor, in the vicinity of the obstacle. This is exactly true for a plane wave ($\mathcal{E}$=const.) and for other sources it becomes more accurate as the distance from the source increases. The position r_0 is an equivalent center of the source. We now insert Eqs. (13) and (14) into Eq. (12), and consider the variations of the exponential factors only. Before writing the resulting equation, it is convenient to introduce the following angles (n is normal to S_I into obstacle):

$$\gamma=\text{angle between } r'-r_0 \text{ and } n;\ \text{thus } \cos\gamma=\frac{\partial|r'-r_0|}{\partial n}, \qquad \beta=\ r'-r_0 \text{ and } i_1;\ \cos\beta=\frac{\partial|r'-r_0|}{\partial u_1},$$
$$\delta=\ r'-r_0 \text{ and } i_2;\ \cos\delta=\frac{\partial|r'-r_0|}{\partial u_2}, \qquad \alpha=\ r'-r \text{ and } n;\ \cos\alpha=\frac{\partial|r'-r|}{\partial n}. \quad (15)$$

Then Eq. (12) becomes

$$E_{(1)}(r)\approx\frac{1}{4\pi}\int_{S_I}\Big\{n\mathcal{E}_n\Big[\Big(1-\frac{\epsilon_a'}{\epsilon_b'}\Big)(-k_a^2\cos^2\alpha+2ik_aG_m\cos\alpha)+\Big(\frac{\epsilon_b'}{\epsilon_a'}-1\Big)(ik_a\cos\alpha+2G_m)$$
$$\times\Big(ik_a\cos\gamma+2G_m\Big[1-\frac{\epsilon_a'}{\epsilon_b'}\Big]\Big)+(k_b^2-k_a^2)\frac{\epsilon_a'}{\epsilon_b'}\Big]+i_1\Big[\frac{i\omega}{c}(\mu_b-\mu_a)\Big(ik_a\mathcal{H}_2\cos\alpha+ik_a\mathcal{H}_2\cos\gamma$$
$$+\frac{i\omega}{c}[\epsilon_b'-\epsilon_a']\mathcal{E}_1+ik_a\mathcal{H}_n\Big(\frac{\mu_a}{\mu_b}-1\Big)\cos\delta\Big)-\Big(\frac{\epsilon_a'}{\epsilon_b'}-1\Big)k_a^2\mathcal{E}_n\cos\alpha\cos\beta+\Big(1-\frac{\epsilon_b'}{\epsilon_a'}\Big)$$
$$\times\Big(-k_a^2\mathcal{E}_n\cos\alpha\cos\beta+2ik_aG_m\mathcal{E}_n\Big(1-\frac{\epsilon_a'}{\epsilon_b'}\Big)\cos\beta\Big)+(k_b^2-k_a^2)\mathcal{E}_1\Big]$$
$$+i_2\Big[-\frac{i\omega}{c}(\mu_b-\mu_a)\Big(ik_a\mathcal{H}_1\cos\alpha+ik_a\mathcal{H}_1\cos\gamma-\frac{i\omega}{c}[\epsilon_b'-\epsilon_a']\mathcal{E}_2+ik_a\mathcal{H}_n\Big[\frac{\mu_a}{\mu_b}-1\Big]\cos\beta\Big)$$
$$-\Big(\frac{\epsilon_a'}{\epsilon_b'}-1\Big)k_a^2\mathcal{E}_n\cos\alpha\cos\delta+\Big(1-\frac{\epsilon_b'}{\epsilon_a'}\Big)\Big(-k_a^2\mathcal{E}_n\cos\gamma\cos\delta+2ik_aG_m\mathcal{E}_n\Big[1-\frac{\epsilon_a'}{\epsilon_b'}\Big]\cos\delta\Big)$$
$$+(k_b^2-k_a^2)\mathcal{E}_2\Big]\Big\}\frac{\exp[ik_a(|r-r'|+|r'-r_0|)]}{|r-r'|}dS. \quad (16)$$

The integral in Eq. (16) can be evaluated approximately by the method of stationary phase. The result is (the factors P and $\mathcal{G}$ are given in Eqs. (19) and (18), respectively):

$$hE_{(1)}(r)=P\cdot\mathcal{G}\cdot\left[\frac{k_a h}{2}\right]\Bigg[n\mathcal{E}_n\Bigg\{\left(1-\frac{\epsilon_a'}{\epsilon_b'}\right)\left(\frac{2iG_m}{k_a}-\cos\alpha\right)+\left(\frac{\epsilon_b'}{\epsilon_a'}-1\right)\left(i+\frac{2G_m}{k_a\cos\alpha}\right)\left(i\cos\gamma+\frac{2G_m}{k_a}\left[1-\frac{\epsilon_a'}{\epsilon_b'}\right]\right)$$
$$+\left(\frac{k_b^2}{k_a^2}-1\right)\frac{\epsilon_a'}{\epsilon_b'\cos\alpha}\Bigg\}+i_1\Bigg\{\frac{i\omega}{k_a c}(\mu_b-\mu_a)\left(i\mathfrak{IC}_2\left[1+\frac{\cos\gamma}{\cos\alpha}\right]+\frac{i\omega}{k_a c\cos\alpha}[\epsilon_b'-\epsilon_a']\mathcal{E}_1\right.$$
$$+i\frac{\cos\delta}{\cos\alpha}\mathfrak{IC}_n\left[\frac{\mu_a}{\mu_b}-1\right]\Bigg)-\left(\frac{\epsilon_a'}{\epsilon_b'}-1\right)\cos\beta\mathcal{E}_n+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\left(-\frac{\cos\gamma}{\cos\alpha}+\frac{2iG_m}{k_a\cos\alpha}\left[1-\frac{\epsilon_a'}{\epsilon_b'}\right]\mathcal{E}_n\cos\beta\right)$$
$$+\left(\frac{k_b^2}{k_a^2}-1\right)\frac{\mathcal{E}_1}{\cos\alpha}\Bigg\}+i_2\Bigg\{-\frac{i\omega}{k_a c}(\mu_b-\mu_a)\left(i\mathfrak{IC}_1\left[1+\frac{\cos\gamma}{\cos\alpha}\right]-\frac{i\omega}{k_a c\cos\alpha}[\epsilon_b'-\epsilon_a']\mathcal{E}_2+i\frac{\cos\beta}{\cos\alpha}\mathfrak{IC}_n\left[\frac{\mu_a}{\mu_b}-1\right]\right)$$
$$\left.-\left(\frac{\epsilon_a'}{\epsilon_b'}-1\right)\cos\delta\mathcal{E}_n+\left(1-\frac{\epsilon_b'}{\epsilon_a'}\right)\left(-\frac{\cos\gamma}{\cos\alpha}+\frac{2iG_m}{k_a\cos\alpha}\left[1-\frac{\epsilon_a'}{\epsilon_b'}\right]\right)\mathcal{E}_n\cos\delta+\left(\frac{k_b^2}{k_a^2}-1\right)\frac{\mathcal{E}_2}{\cos\alpha}\Bigg\}\right]. \quad (17)$$

5. PROPERTIES OF THE SOLUTION

In Eq. (17) the first-order scattered (reflected, transmitted, or diffracted) field has been written as a product of three factors, each of which has special physical significance. It is interesting to observe that the solution of the corresponding acoustic problem has a similar form, the first two factors being identical with those found here.[3] We shall now discuss each of the factors in turn.

Geometrical Factor

The factor $\mathcal{G}$ in Eq. (17) depends upon geometrical quantities only and is given by

$$\mathcal{G}=\Bigg\{1+D_2\left(\frac{1}{D_1}\left[\frac{2-\sin^2\gamma}{\cos^2\alpha}-\tan^2\alpha+2\tan^2\alpha\sin^2\alpha\sin^2(\theta-\varphi)\right]-(\cos\alpha+\cos\gamma)[2G_m+G_{||}(\theta)\tan^2\alpha]\right)$$
$$+\left(\frac{\cos\gamma}{\cos\alpha}\frac{D_2}{D_1}\right)^2+D_2^2G_g\left(1+\frac{\cos\gamma}{\cos\alpha}\right)^2-\frac{D_2}{D_1}\frac{\cos\alpha+\cos\gamma}{\cos^2\alpha}(2G_m\cos^2\gamma+G_{||}(\varphi)\sin^2\gamma)\Bigg\}^{-\frac{1}{2}}\frac{\cos\alpha}{|\cos\alpha|}. \quad (18)$$

The various symbols in Eq. (18) are defined as follows, at the point on S_I where the phase in Eq. (16) is stationary: θ=angle between u_1 and the plane containing the normal and r_1, φ=angle between u_1 and the plane containing the normal and r_2, G_m=mean curvature, G_g=Gaussian curvature, $G_{||}(\theta)$=curvature of the normal section of surface S determined by the plane making angle θ with u_1, $D_1=|r'-r_0|$, $D_2=|r'-r|$.

We say that this factor accounts for the geometrical spreading or focusing of the scattered field because it also occurs in the purely geometrical optical solution of the reflection problem. In fact, the ratio of the field amplitude reflected from an interface to the incident amplitude has been computed by geometrical optics[2,3] and found to be precisely the above quantity $\mathcal{G}$ in directions of specular reflection ($\alpha=\gamma$, $\theta=\varphi$) and zero in other directions (except for a reflectivity factor.) This factor has been investigated in references 2 and 3 and shown to yield the positions of caustics, foci, point images, focal length, and the mirror law. $\mathcal{G}$ becomes infinite on caustics and at foci, but a more careful evaluation of the integral in Eq. (16) at such places, employing Airy integrals instead of Fresnel integrals, yields finite but large values of the field.

Phase Factor

The factor P in Eq. (17) depends upon the incident wave-length and the obstacle dimensions, as well as the distances from source and observation point to the surface. It is given by

$$P=\frac{e^{iM'}}{2}[F_\pm(\xi_0'')-F_\pm(\xi_1'')][F_\pm(\eta_0'')-F_\pm(\eta_1'')]. \quad (19)$$

The symbols in Eq. (19) are defined as follows (see reference 3 for further details):

$F_\pm(x)=\int_0^x \exp[\pm i(\pi/2)u^2]du$=Fresnel integral; plus or minus sign chosen in accordance with the rule given in reference 3.

$$M'=k_a(D_1+D_2)+\cdots.$$

P oscillates as the wave-length varies, with its absolute value ranging between zero and about two. It generally has a maximum in the direction of specular reflection, and approaches one in this direction, and zero in all other directions, as the wave-length approaches

zero, thus yielding the optical "law of reflection." In addition it yields a 180° phase change on crossing a point image or focus. Reflection from an interface has been treated by the Kirchhoff method[3,4] and the reflected field found to equal the incident field multiplied by $\mathcal{G}P$. The arguments of the Fresnel integrals in Eq. (19) are given by

$$\xi_0''=\frac{\sqrt{2}\xi_0'}{\left(\frac{\pi}{|A'|}\right)^{\frac{1}{2}}},\quad \eta_0''=\frac{\sqrt{2}\eta_0'}{\left(\frac{\pi}{|C'|}\right)^{\frac{1}{2}}} \tag{20}$$

with similar expressions for ξ_1'', η_1''. The quantities $\xi_0'\xi_1'\eta_0'\eta_1'$ determine the boundaries of a rectangle equivalent to the "illuminated" part of the surface, referred to the point of specular reflection as origin. The denominators in Eq. (20) are just the half-dimensions of the first Fresnel zone on the surface around the specular reflection point. If the dimensions of the surface are large compared to those of the first Fresnel zone, the Fresnel integrals have their asymptotic values and the phase factor has absolute value one (provided medium a has zero conductivity, i.e., k_a is real). This is usually the case unless the dimensions of the surface are small compared to a wave-length or it is cylindrical or plane.

When the surface is small compared to the first Fresnel zone the arguments of the Fresnel integrals are small, and these integrals may be approximated by the first terms in their Taylor expansions, which yield $F_{\pm}(x)\approx x$. Then

$$P=e^{iM'}\frac{(\xi_0'-\xi_1')(\eta_0'-\eta_1')}{\pi/(|A'C'|)^{\frac{1}{2}}}. \tag{21}$$

From Eq. (21) we see that, apart from the exponential factor, P is equal to the ratio of the area of the obstacle surface to the area of the first Fresnel zone.

Since both A' and C' have the factor k_a, which is equal to $2\pi/\lambda_a$, where λ_a is the wave-length in medium (a), if the conductivity of medium (a) is zero, P is inversely proportional to λ_a. The reflectivity factor also has the factor $k_a h$, and hence one sees that the scattered field is proportional to the obstacle volume (=area$\times h$) and inversely proportional to λ_a^2. This is the Rayleigh scattering law for electromagnetic waves. In particular, it is valid if $\lambda_a \gg$ obstacle dimensions, since then the first Fresnel zone is larger than the obstacle.

Reflectivity Factor

The scattered field is expressed in Eq. (17) in terms of components in the n, i_1, i_2 directions, where n, i_1, i_2 are mutually orthogonal unit vectors at the point of stationary phase on the surface S. Each of these components is the product of the geometrical and phase factors and a third factor which is called the reflectivity factor. Since this factor is not the same for all components, it is appropriate to speak of three reflectivity factors, one for each component. The factors are the coefficients of n, i_1, and i_2 in Eq. (17).

The reflectivity factors are linear and homogeneous in the components of the incident field. They depend upon the physical properties of both media, the thickness and curvature of the shell, and the angles of incidence and reflection. Thus it can be seen that these factors take account of the propagation within the shell. From an examination of the form of these factors it is seen that they are proportional to the amplitude of the incident field, and increase in absolute value as $|\epsilon_a'-\epsilon_b'|$ and/or $|\mu_a-\mu_b|$ increase. Moreover, they are proportional to $k_a h$ which, in the case when medium (a) has zero conductivity, is $2\pi h/\lambda_a$. Thus the scattered field increases with the ratio of the thickness to the wave-length. This inverse dependence upon λ_a, in conjunction with the phase factor, leads to the electromagnetic analog of the Rayleigh scattering law, as has been pointed out above. When reflection and diffraction problems are treated by the Kirchhoff method no reflectivity factor is obtained, and hence the Rayleigh scattering law cannot be deduced.

If the reflectivity factors are divided respectively by the corresponding field components, they agree with the exact coefficients[9] for the reflection of plane waves from a thin, infinite, plane plate, except for additional terms proportional to the mean curvature of the surface. The polarization effects are therefore practically the same as for the plane case above. If Eq. (17) is specialized to the case of plane waves incident on a plane shell, the geometrical factor and the phase factor both become unity, and the reflected and transmitted fields agree exactly with those obtained by Luneberg[9] when his results are applied to a thin plate.

The presence of the quantity $\cos\alpha$ in the denominator of some terms of the reflectivity factors appears to indicate that the scattered field becomes infinite as $\alpha\rightarrow\pi/2$. In this case, however, the method used to evaluate the integral is not valid (see above).

CONCLUSION

The solution seems to yield a rather satisfactory quantitative as well as qualitative understanding of reflection and transmission by thin shells, as well as agreement with various special results previously obtained by other methods. For thick shells additional terms in the series solution would probably be needed. Since the calculations required would be quite complicated, an alternative method which was useful in acoustics seems preferable. We merely assume that the reflected field still contains three factors, the geometrical and phase factors being the same as those given here and the reflectivity factor being that obtained for a thick

[9] R. K. Luneberg, "The propagation of electromagnetic plane waves in plane parallel layers"; Research Report No. 172-3 to the Watson Laboratories; New York University, Washington Square College Mathematics Research Group (June, 1947).

plate. The physical interpretation of the various factors makes this procedure at least plausible. Calculations based on these assumptions seemed to agree with experimental results in acoustics.

A further test of the result of this paper is provided by an exact solution for the reflection and transmission of spherical waves by a spherical shell with a dipole source at its center.[10] This solution was specialized to the case of a thin shell and compared to the result of the present paper when applied to this case. The agreement between the two solutions was satisfactory for all components except E_r, the radial component of electric field, and except near the source (see reference 10 for further details).

[10] J. B. Keller and H. B. Keller, J. App. Phys. 20, 393-396 1949).

ACKNOWLEDGMENT

I wish to express my appreciation to Professor Henry Primakoff for his excellent instruction, advice and direction in the solution of the corresponding acoustic problem.

A Technique for Taking Motion Pictures of Electron Microscope Images*

LUTHER E. PREUSS AND JOHN H. L. WATSON
Edsel B. Ford Institute for Medical Research, Detroit 2, Michigan
(Received April 12, 1950)

Lately, for the first time, a technique for taking motion pictures of electron microscope images has been developed which is proving useful in the study of some of the effects of electrons upon matter and particularly upon colloidal crystals. Resolution of 150 to 300A is achieved and is suitable for a host of practical electron microscopical problems. Pictures are taken from outside the instrument, utilizing the radiation emitted by the fluorescent screen. Some instrument adaptation is required for adequate illumination, and precautions must be exercised against toxic x-ray dosages.

THE feasibility of taking motion pictures of electron microscope images has been considered on occasion, but a lack of informative, moving subjects made the project impractical. Lately, the technique has been developed successfully in these laboratories for studying the effects of electrons, and indirectly of heat upon colloidal crystals. The effects of electrons upon a variety of other crystalline and non-crystalline materials are also being studied in the same way, and many specimen changes are observed in a much more positive manner than heretofore. Some reactions which escape detection entirely or which occur too rapidly for satisfactory observation to be made at all by the orthodox procedures are registered on motion picture films. Pronounced reactions have been recorded in one frame intervals in as short a time as 0.03 sec.

The motion pictures are taken by photographing the image on the fluorescent screen with the camera set up outside the microscope. Figure 1 shows the relative positions of camera and microscope with the final viewing screen tilted so as to have its plane normal to the optical axis of the camera and to give a maximum field of view. The angle is about 30 degrees. Some distortion is introduced by the tilt, and this has to be kept in mind when interpreting the images, but it is not found to be bothersome. The camera is a 16-mm Cine Kodak Special with a 4-in., $f/2.7$ telephoto lens mounted upon a conventional tripod. This tripod is a source of error due to vibration and a more rigid camera support is recommended. Practical operating speeds are from 8 to 32 frames per second, depending upon screen intensity. Above 32 frames per second the contrast deteriorates. Super XX negative panchromatic safety film has been found to produce the best contrast and Super XX and Kinolux positive emulsions are used with moderate success. The complete processing time with *D*-19, exclusive of drying, is at least two hours per 100-ft. roll of 16-mm negative. To reduce the amount of manual work required in the processing and to shorten the time, a motor-driven tank has been built, Fig. 2.

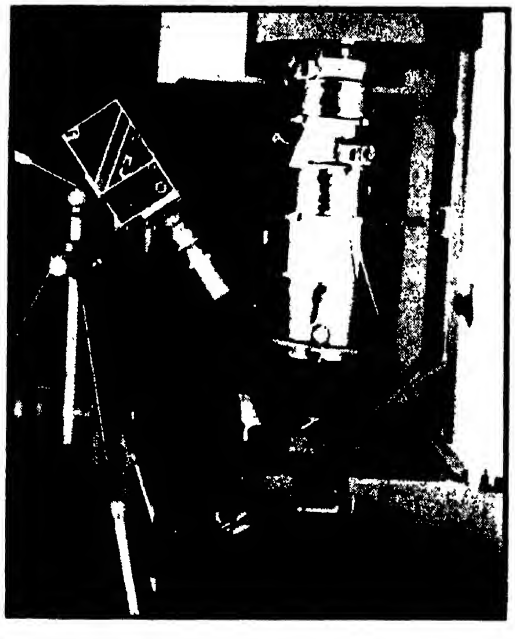


FIG. 1. Experimental arrangement of camera and microscope, showing angle of final viewing screen.

The intensity of the light emitted from the final viewing screen has to be of a relatively high order. A self-biasing electron gun[1] is the first requirement, both for sufficient intensity and for noticeable effects of bombardment upon samples. It is necessary to replace the fluorescent materials on the screens periodically because they become contaminated and lose their brilliance. Since much of the work has to be done in the lowest portion of the linear section of the Hurter-Driffield curve, the image brightness has to be increased to its maximum without provoking premature or undesirable specimen changes. In general, although the contrast would be desirable, no objective aperture can be used except with materials which react very slowly under the beam. Objective apertures were used in taking moving micrographs of colloidal crystals[2] of β-FeOOH under bombardment, but none were usable with NaCl or WO_3 crystals where the desired effects occurred more rapidly.

In order to increase the possible intensity still further, the diameter of the condenser aperture is opened from 25 to 50 mils. This, plus the tilting of the final viewing

* Presented in part before the annual meeting of the Electron Microscope Society of America, Washington, D. C. (1949).

[1] James Hillier and R. F. Baker, J. App. Phys. 17, 12 (1946).
[2] John H. L. Watson, J. App. Phys. 19, 713 (1948).

screen, results in the production of a considerable amount of x-radiation and makes it imperative that the instrument be surveyed. Some remarkably high dosage rates were measured for the experimental conditions. Figure 3 shows rough isodose curves at the position of the final viewing screen.

The electron microscope, which has inherent in it both a source of heat from electron bombardment in vacuum and a means of photography from the same radiation, offers a convenient means of conducting studies of solid-liquid transformations. The application of the moving picture technique provides a method for investigating the crystalline properties of such transformations by observation of Bragg electron reflections in specimens under bombardment. In normal electron microscopy Bragg reflections are seen to best advantage in dark field,[3] but in bright or dark field are most easily and definitely detectable when caused to move by variations of objective lens current. When the object itself is crystallizing or melting, or is in some other unstable condition, the images of the reflections move on the screen and are detectable without an objective current variation. They may be recorded effectively, therefore, by the motion picture technique. From direct observation, the bombarded samples are seen to emit a profusion of such reflections which originate from what appear from characteristics of shape alone to be melted areas as well as from those regions which are judged to be composed of well-formed crystals. Presumably, some of the ostensibly melted areas must often retain some oriented forms.

The motion picture technique for electron microscopes should also be useful for studying samples which are under the influence of factors other than electron bombardment in the microscope. For example, observations could be recorded to advantage upon film, if heat effects were studied, where the heating was done by a direct furnace source rather than by the beam. Ideal objects for study would also be provided during the deposition of thin metal films within the vacuum of the electron microscope, as reported[4] by Sennett of the Toronto group. The method is promising for a detailed study of the effects of electron bombardment upon bacteriological and biological specimens, and whenever a method is

[3] C. E. Hall, J. App. Phys. 19, 198 (1948).

[4] R. S. Sennett and T. A. McLauchlan, University of Toronto, unpublished research, presented before the annual meeting of the Electron Microscope Society, Washington, D. C. (1949).

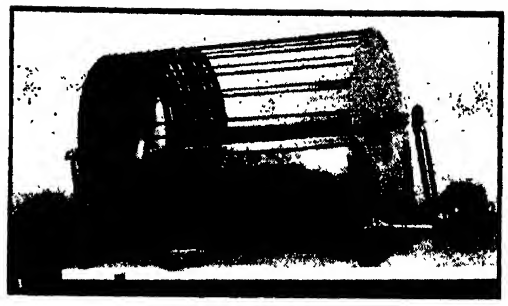


FIG. 2. Motor-driven, processing tank.

devised for studying live or wet specimens satisfactorily by electron microscopy the advantages of recording the observations on movie film will be obvious.

There are a number of other incidental advantages to be gained from taking motion pictures of electron microscope images. Non-electron microscopists, who are not as familiar with the appearance of images in the microscope but who are collaborating on a problem, find

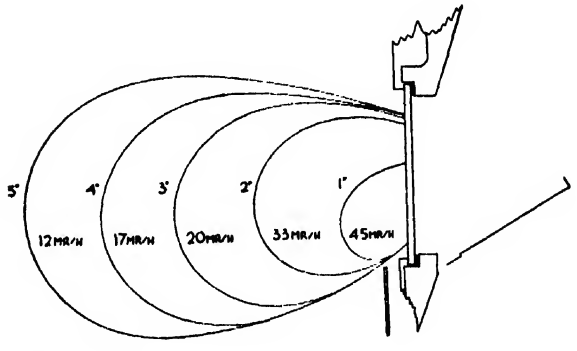


FIG. 3. Rough isodose curves drawn in location of tilted final viewing screen, with a 50-mil condenser aperture and sufficient intensity for taking motion pictures. Values would be much higher from intermediate screen.

the films very instructive. The same is true in lectures on electron microscopy where an audience is given an improved impression if a movie sequence is shown as illustration, even when the motions recorded are themselves not particularly informative for new things. The ability to reverse or retard a film is very useful, especially when observing the extremely fast reactions which occur during the earliest stages of a bombardment series.

Motion Picture Studies of Electron Bombardment of Colloidal Crystals

JOHN H. L. WATSON AND LUTHER E. PREUSS
Edsel B. Ford Institute for Medical Research, Detroit 2, Michigan

(Received April 12, 1950)

Phenomena not seen or recorded by orthodox methods are observed on the large number of micrographs which are taken during a motion picture experiment, and reaction times as short as 0.03 sec. can be calculated. The technique is useful for investigating solid-liquid-solid changes both by direct observation and by recording Bragg reflections. In most cases a preliminary effect of electrons upon crystals is to drive off water. In the case of sodium chloride crystals a residual envelope is left. Inside this, small particles, both solid and liquid, are observed to move rapidly due, it is believed, to a combination of Brownian movement and convection. These gradually evaporate and diffuse through the wall. The possibility is discussed that the residual envelope is formed by contamination. Examples are also taken from bombarded, colloidal crystals of tungsten oxide.

GERMAN workers realized at an early date the advantages to be gained in electron microscopy by the use of a battery-biased source, and described some of the effects upon their specimens of the intense beam from such a gun. The major emphasis was placed upon the subject in this country after Hillier and Baker[1] described the self-biasing cathode. Several reports[2–4] have been published since which discuss from stills the effects of electrons upon matter as observed in electron microscopes employing self-biasing sources, and observations of these reactions are often used now to clarify electron microscopic sample interpretation.

Several colloidal crystal systems have been studied by application of the motion picture technique[5] for their reactions under electron radiation. The first of these involves crystals of sodium chloride.[6] In Fig. 1 several stills are shown which are typical of a movie sequence. It will be seen that the contrast of the "moving micrographs" is good and that the resolution and magnifications are useful, if not completely comparable with the best to be expected from the more orthodox methods of electron microscopy. Figure 1A shows an early stage of the bombardment series as the water is being driven off and sections of the crystal are becoming transparent to electrons. Before this the whole crystal was quite opaque. Figures 1 (B–F) are all progressive, intermediate appearances of the same crystal as it changes toward the residual envelope[6] condition which is its final state, Fig. 1G.

The fine particles, which are seen to form within the envelope, move very rapidly in a random fashion under bombardment. This motion slows or ceases whenever the intensity is diminished. From higher resolution micrographs taken in the orthodox manner, Fig. 2, these particles are often seen to be rounded, but in many cases are small cubes. The rounded particles are apparently molten and are frequently noticed to coalesce with others which they meet. The cubes are presumed to be cooler, solid crystals. All the particles continue to decrease in size and contrast until eventually they evaporate completely, diffuse through the envelope as vapor and disappear.

The motion of the particles is probably a combination of convection and Brownian movement of the particles within the confined space of the envelope under the influence of the warm, vaporous material which fills it. That it is filled with such a material under pressure is shown by the facts (1) that the sides of the envelope will often bulge outward and even break if heat is applied too rapidly and (2) that such a break results in the spewing out of vaporous substance from the envelope which deposits as solid particles on the cool, surrounding substrate. These effects are demonstrated very graphically in the moving pictures. It has been thought that Brownian motion would not be studied to advantage in an electron microscope due to the extremely large translations which would be involved at the high magnifications on the very limited viewing screen. Here, however, the motion is constrained within narrow limits by the residual envelope. Figure 3 from another movie sequence shows a ruptured envelope.

The residual envelope has been conjectured[6] to originate from contamination[7] but if so, the contamination must be much more quickly and effectively deposited upon some materials than upon others, for in these studies only sodium chloride has been found to leave such envelopes. This observation agrees with the results of Burton *et al.* who classed sodium chloride along with sodium chlorate, potassium bromide, and iodide as substances which behaved in this way. If contamination explains the occurrence of residual envelopes, this specificity for certain substances is in disagreement with the qualitative observations by Hillier[8] that the rate of deposition was the same for all substances. On the other hand, if Hillier's observation is correct, other explanations must be looked for to explain the presence of

[1] James Hillier and R. F. Baker, J. App. Phys. **17**, 12 (1946).
[2] John H. L. Watson, J. App. Phys. **19**, 713 (1948).
[3] Heller, Wojtowicz, and Watson, J. Chem. Phys. **16**, 998 (1948).
[4] F. A. Hamm, and Earl Van Norman, J. App. Phys. **19**, 1097 (1948).
[5] Luther E. Preuss and John H. L. Watson, J. App. Phys. **21**, 902 (1950), preceding paper.
[6] Burton, Sennett, and Ellis, Nature **160**, 565 (1947).
[7] John H. L. Watson, J. App. Phys. **18**, 153 (1947).
[8] James Hillier, J. App. Phys. **19**, 226 (1948).

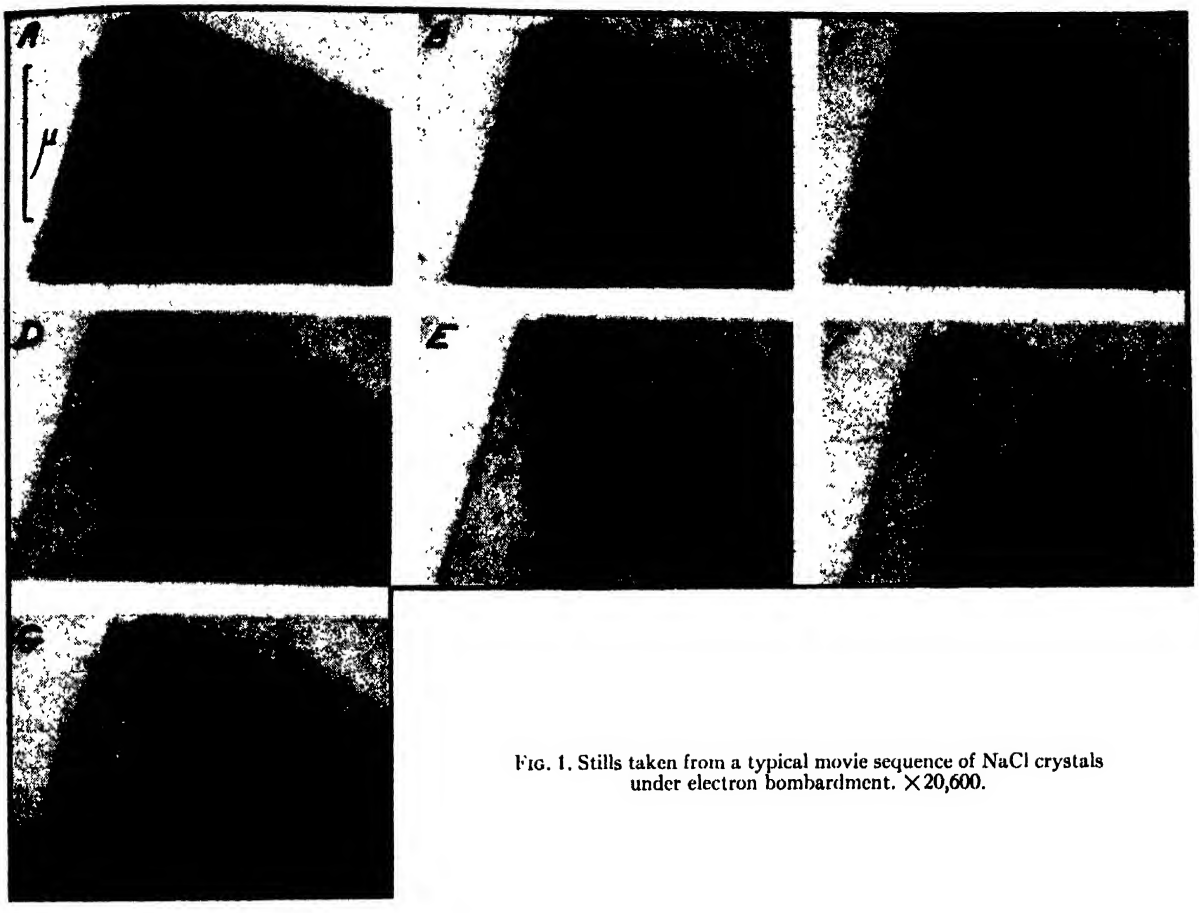


FIG. 1. Stills taken from a typical movie sequence of NaCl crystals under electron bombardment. ×20,600.

residual envelopes. The envelope is always present with sodium chloride regardless of the speed of evaporation, which often occurs instantaneously. It may be significant that those salts which are water soluble tend to leave residual envelopes while the insoluble ones do not. It is interesting to hypothesize that the residual layer is left as a result of trace impurities which concentrate at the original surface during the evaporation of the interior. Salt crystals deposited from saline, organic suspensions yield particularly striking envelopes, Fig. 4, a fact which may indicate that organic impurities are polymerized or condensed at the crystal surface in a radiation-chemical process under the action of the electron beam.

The three-dimensional nature of the residue as a replica of the original crystal has been demonstrated by shadow-casting them after evaporation and by tilting them on broken Formvar films and recording their appearance in movie film. As a replica they give information concerning the surface of the original crystals.

The large number of micrographs taken rapidly by the moving picture technique has resulted in several examples being recorded of a transient appearance of an inherent mosaic structure in sodium chloride crystals, in support of the discussion by Ewald and Renninger[9] that a mosaic exists. The phenomenon was observed to form and disappear quickly, and its presence was detected in only one or two stills of any sequence in which it ap-

FIG. 2. Still taken by orthodox, electron microscope procedure of NaCl crystal under electron bombardment. Reversed print. ×12,000.

[9] P. P. Ewald and M. Renninger, International Conference on Physics Papers and Discussions (Physical Society, 1935), Vol. 11, pp. 57–61.

FIG. 3. Still taken from a typical movie sequence showing a ruptured envelope at arrow. ×26,000.

peared. Its visibility was therefore a short-lived affair which would not be micrographed as successfully without the use of the moving picture methods. One still is shown in Fig. 5 as an illustration. It is an interior rather

FIG. 4. Electron micrograph of evaporated salt crystals in a bacterial preparation. Reversed print ×3600.

than a surface effect, since it disappears under continued bombardment.

The bombardment effects of electrons upon tungsten oxides (WO_3) have been reported earlier,[1,2] but are more

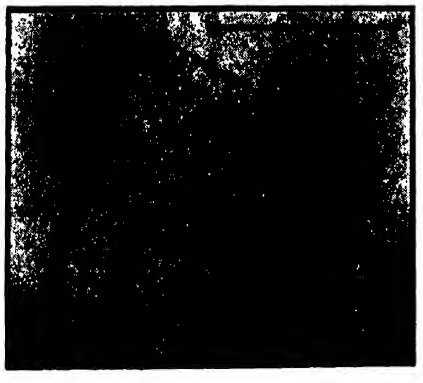


FIG. 5. Still taken from a typical movie sequence, showing the mosaic structure in an evaporating NaCl crystal. ×26,000.

completely explainable from observations made on film. By slowing down the effects considerably it has been confirmed that the mosaic pattern of rods in fresh, early samples is inherent. The initial effect of bombardment is to render these visible but not to create them. Apparently, the heat from the incident beam drives off water of crystallization and some free passage of electrons results. The initial stages of bombardment of these and of most other crystals passes extremely rapidly, depending upon the intensity, and these stages are the most productive in moving picture studies. The effects of melting and recrystallization into other forms are secondary and enter later.

FIG. 6. A typical sequence of frames taken of the effects of electron bombardment upon WO_3. The sequence was taken at 16 frames per second and covered a total time interval of $\frac{1}{4}$ sec. ×7200.

Figure 6 shows a typical sequence of frames taken at an original magnification of ×7200 in the microscope and reproduced here at the same magnification. Since these were taken at 16 frames per second, the observed reactions occurred within a very short interval. The actual exposure time per frame for this example was 1/30 sec., and the whole sequence represents a time interval of $\frac{3}{4}$ sec. In the first one or two frames little if any reaction has occurred although some of the inherent structure is seen to be giving rise to Bragg reflections which appear as a straight, intense line over the crystal surface. Later the inherent mosaic has become visible over the whole body of the somatoid. In the final frames secondary structures due to melting have been introduced at the tip of the crystal.

There is no suggestion of a residual envelope when tungsten oxide is bombarded. Melting is recognizable from the round nature of the melted particles, and recrystallization is interpreted from production of moving Bragg reflections and from a return to the straight-edged morphology. Areas presumed to be melted often give rise to electron reflections. Melted particles are also seen to "bubble" in the moving micrographs as if boiling under the beam.

Moving pictures of carbon black contaminating under the beam demonstrated the particle growth very graphically, but added no new knowledge of the phenomenon.

The Emission of Radiation from Nitric Oxide: Approximate Calculations*

L. E. Benitez** and S. S. Penner
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California
(Received March 27, 1950)

Emissivity calculations have been carried out for nitric oxide at temperatures from 300°K to 3000°K. The results may be used, as a first approximation, at elevated total pressures.

NITRIC oxide[1] is one of the diatomic molecules which may contribute to radiant heat transfer in high pressure combustion chambers. In particular, nitric oxide at elevated temperatures is encountered in many solid-fuel and liquid-fuel rocket motors. It is, therefore, of obvious practical importance to determine the emissivity of nitric oxide as a function of temperature. According to the procedure for making approximate radiant heat-transfer calculations outlined previously,[2] it is necessary to calculate first an effective band width lations are then used to construct a plot of limiting emissivities for the fundamental and the first overtone of NO as a function of temperature. Finally, emissivity calculations are made by using the limiting emissivities in conjunction with the values of the integrated absorption for the fundamental and the first overtone, estimated from available experimental absorption measurements of NO.

The results of effective band-width calculations, using the procedure described previously[2] and the spectroscopic constants given by Mayer and Mayer,[3] are summarized in Tables I and II for the fundamental and the first overtone of NO, respectively, for two different values of $j_{max.}$ at each temperature. It should be noted that the presence of a Q branch in the infra-red spectrum of NO does not affect the effective band-width calculations made according to our approximate procedure. For values of the limiting rotational quantum number $j_{max.}$ exceeding the value corresponding to the position

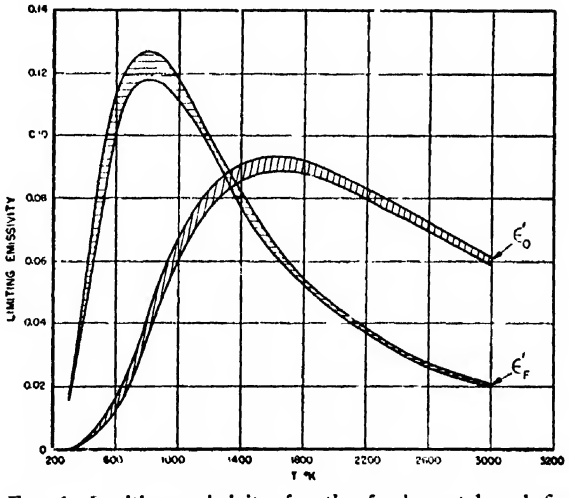


Fig. 1. Limiting emissivity for the fundamental and first overtone of NO, as a function of temperature, for two sets of values of $j_{max.}$.

* This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. W-04-200-ord-455, sponsored by the U. S. Army Ordnance Department.

** Lieutenant, U. S. Navy. Present address: Massachusetts Institute of Technology, Cambridge, Massachusetts.

[1] Further details concerning radiant heat-transfer calculations on nitric oxide may be found in a thesis submitted by Lt. L. E. Benitez in partial fulfillment of requirements for the degree of Aeronautical Engineer, California Institute of Technology, June 1949.

[2] S. S. Penner, J. App. Phys. 21, 685 (1950).

[3] J. E. Mayer and M. Goeppert-Mayer, *Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1940), p. 469. The numerical results in the present report are practically unchanged if the spectroscopic constants given by R. H. Gillette and E. H. Eyster, Phys. Rev. 56, 1113 (1939), are used.

TABLE I. Effective band width of the fundamental of NO as a function of temperature.

T °K	$j_{max.}$	$\nu_{max.}$ cm^{-1}	$\nu_{min.}$ cm^{-1}	$\Delta\nu_{n\to n+1}$ cm^{-1}
300	30	1962	1761	201
	33	1968	1748	220
1000	54	2003	1646	357
	58	2007	1625	382
1500	65	2013	1587	426
	69	2015	1566	449
2000	77	2018	1519	499
	81	2019	1496	523
2500	88	2019	1454	565
	91	2019	1436	583
3000	100	2019	1379	640
	106	2019	1341	678

TABLE II. Effective band width of the first overtone of NO as a function of temperature.

T °K	$j_{max.}$	$\nu_{max.}$ cm^{-1}	$\nu_{min.}$ cm^{-1}	$\Delta\nu_{n\to n+2}$ cm^{-1}
300	30	3793	3594	199
	33	3796	3577	219
1000	54	3801	3442	359
	58	3801	3412	389
1500	65	3801	3360	441
	69	3801	3328	473
2000	77	3801	3260	541
	81	3801	3226	575
2500	88	3801	3162	639
	91	3801	3134	667
3000	100	3801	3046	755
	106	3801	2985	816

of a band head, the limiting frequency has been set equal to the frequency at the band head.[2]

From the data listed in Tables I and II the limiting emissivities have been calculated. The limiting emissivities ϵ_F' and ϵ_0' for the fundamental and the first overtone, respectively, are defined by the relations

$$\epsilon_F' = \int_{\Delta\nu_{n\to n+1}} \rho(\nu)d\nu \Big/ \int_0^\infty \rho(\nu)d\nu,$$

$$\epsilon_0' = \int_{\Delta\nu_{n\to n+2}} \rho(\nu)d\nu \Big/ \int_0^\infty \rho(\nu)d\nu,$$

where $\rho(\nu)$ is the energy density of radiation emitted by a blackbody at a given temperature. The limiting emissivities for the two sets of $j_{max.}$ values are plotted as a function of temperature in Fig. 1.

In order to utilize the plot of limiting emissivities shown in Fig. 1 for radiant heat-transfer calculations, it is necessary to estimate the integrated absorption for the vibration-rotation bands.[2] Reliable numerical values do not appear to be available for NO although qualitative results have been reported by Gillette and Eyster[3] and by Rideal *et al.*[4] Experimental measurements of apparent absorption coefficients of NO and CO at small total pressures with an instrument of low spectral resolution have led to similar experimental results for the integrated absorption for these two gases.[1] On the basis of these studies it is suggested that the emissivity ϵ' be determined by use of the relations

$$\left.\begin{aligned}
\epsilon' &= [1-\exp(-\bar{k}_F pl)]\epsilon_F' \\
&\quad +[1-\exp(-k_0 pl)]\epsilon_0', \\
k_F(T) &= \frac{201}{\Delta\nu_{n\to n+1}} \times \frac{300}{T} \times 116 \\
&= \frac{7.0\times 10^6}{T\Delta\nu_{n\to n+1}}, \text{ (cm-m-atmos.)}^{-1}, \\
k_0(T) &= \frac{200}{\Delta\nu_{n\to n+2}} \times \frac{300}{T} \times 0.88 \\
&= \frac{5.3\times 10^4}{T\Delta\nu_{n\to n+2}}, \text{ (cm-m-atmos.)}^{-1},
\end{aligned}\right\} \quad (1)$$

where the following numerical values have been used at 300°K:

$$\Delta\nu_{n\to n+1} = 201 \text{ cm}^{-1},$$

$$\Delta\nu_{n\to n+2} = 200 \text{ cm}^{-1},$$

$$\bar{k}_F = 116 \text{ (m-atmos.)}^{-1},$$

$$k_0 = 0.88 \text{ (m-atmos.)}^{-1}.$$

The numerical values of $\bar{k}_F$ and k_0 should be regarded as rough approximations subject to revision on completion of current measurements of integrated absorption.

[4] Snow, Rawlins, and Rideal, Proc. Roy. Soc. (London) 124A, 453 (1929).

Generalized Microscopy and the Two-Wave-Length Microscope

M. J. Buerger
Crystallographic Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received April 27, 1950)

In the ordinary simple microscope, the same light is permitted to flow through the entire system. A more general microscope can be devised in which light of one wave-length is permitted to flow as far as the diffraction image of the object, then light of a second wave-length is substituted at this plane, and continues to flow through the rest of the optical system. It is shown that for such a generalized microscope, the magnification depends not only on the ordinary image-to-object distance ratio, but also on the ratio of wave-lengths used. Utilizing x-rays and visible light, this second factor is about 10^4 diameters, and the whole magnification is of the order of 3×10^6. This is enough to permit one to see an atom. Apparatus has been built which realizes this theory. A most important characteristic of any such system is that a loss of phase occurs at the first diffraction image where the wave-length substitution takes place. This phase must be supplied by the apparatus. It is accomplished with the aid of phase shifters prepared from a large uniform mica cleavage fragment. The apparatus and the results achieved with it are described.

THE generalized theory of microscopy discussed here was developed by the writer in 1939, in connection with the theory of optically reciprocal gratings,[1] and it was presented at the January 10, 1941 meeting of the New York Academy of Sciences, at the October 20, 1942 meeting of the Rochester Section of the American Optical Society, and on a number of subsequent occasions. The subject matter is, therefore, not entirely new. But what makes it timely and interesting is that a technical development which was in a successful experimental stage in 1941 has now been perfected, and therefore new and very interesting results have recently been achieved with this instrument. The gap in development was due to the intervention of the war, and consequent diversion of interests.

The two-wave-length microscope is a device for magnifying an image in two stages, using a different wave-length in each stage. The magnification then depends not only on the usual ratio of image-to-object distances, but also on the ratio of the wave-lengths used. To understand this action, consider first an ordinary optical system employing the same wave-length of light throughout, shown in Fig. 1. This consists of two lenses L_1 and L_2, of focal lengths D_1 and D_2, respectively, and separated by a distance D_1+D_2. An illuminated object is placed at the left focus of the first (left) lens, and its image is formed at the right focus of the second (right) lens. The dotted lines indicate the geometrical optics of image formation.

Now, this simple process can also be regarded as two stages of diffraction in series. The first lens, with the object at its focus, collects parallel rays emanating from the object and focuses these at its right focal plane. This is the diffraction image of the object. Then those rays emanating from this diffraction image which are parallel are focused by the second lens to become the final image. Thus the final image is also the *diffraction of the diffraction* of the object. The focusing of the diffraction in these two stages by the two lenses is indicated in Fig. 1 by the solid lines.

If the object is periodic, then its diffraction image is in the form of a collection of discrete spectra. Let the periodicity of the object be defined by the lattice of the object. Then,[1] as will be shown, the diffraction image is the weighted reciprocal lattice[2] of the object. The diffraction image of this reciprocal lattice is the final image. To generalize this statement for non-periodic objects, one can restate it in the language of Fourier transforms. The diffraction of the object is its Fourier transform. The diffraction of the diffraction image is therefore the *transform of the transform* of the object. It is well known that this is the enantiomorphous equivalent of the object.

Consider the magnification of the image in terms of this double diffraction process. It simplifies the treatment to carry through the discussion of magnification for a periodic object, thus avoiding the explicit use of Fourier transforms. Let the translations t_1 and t_2 define the periodicity of the pattern of the object in its own plane. Then the wavelets scattered by the pattern units of the object are alternatively in phase and out of phase in a plane at right angles to a set of parallel rows, Fig. 2. If the spacing of such rows is d, then the condition that the wavelets are in phase (Fig. 3) is

$$\sin\delta = n\lambda_1/d. \qquad (1)$$

These re-enforcing waves are focused on the plane of the diffraction image at a point whose distance, t_n^*, from

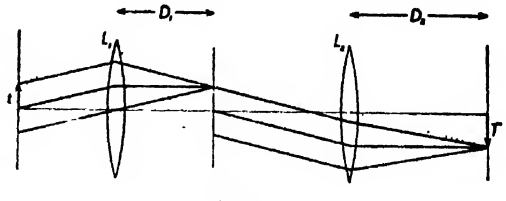


FIG. 1.

[1] M. J. Buerger, "Optically reciprocal gratings and their application to the synthesis of Fourier series," Proc. Nat. Acad. Sci. 27, 117–124 (1941).

[2] M. J. Buerger, *X-Ray Crystallography* (John Wiley and Sons, Inc., New York, 1942), pp. 107–127.

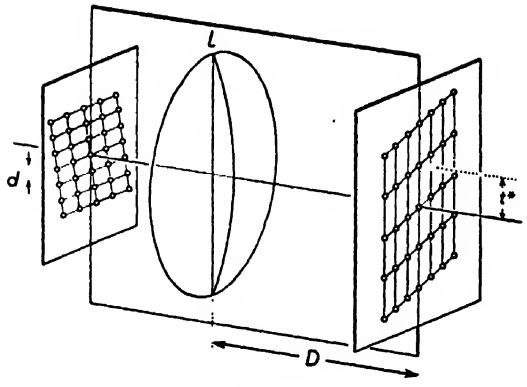


FIG. 2.

the optic axis of the system (Fig. 3) is given by

$$\tan\delta = t_n^*/D_1. \tag{2}$$

Combining (1) and (2),

$$t_n^* = D_1 \tan \sin^{-1}(n\lambda_1/d). \tag{3}$$

Provided that δ is small (i.e., that discussion is limited to spectra comparatively near to the optic axis), a close approximation is

$$t_n^* = n(D_1\lambda_1)/d. \tag{4}$$

Thus, the collection of the n spectra constitutes a row of points whose interval is

$$t^* = (D_1\lambda_1)/d. \tag{5}$$

Such a row of spectra occurs normal to *any* set of rows in the object Fig. 2, i.e., they occur parallel to the d_{hk} of any set of rows of crystallographic index (hk) and the law of the interval t_{hk}^* in the spectrum is

$$t_{hk}^* = K/d_{hk}, \tag{6}$$

where K is an experimental constant consisting of the product $(D_1\lambda_1)$.

A collection of points displaying this kind of reciprocity with respect to the spacings of all the rows of a periodic object is known to be a lattice,[2] and in fact it is the reciprocal lattice of the object. Thus the diffraction of a periodic pattern is a set of discrete spectra at points of the reciprocal lattice of the object.

But this reciprocity is deeper than merely the reciprocity in the dimensions of t^* and d; the reciprocity is truly reciprocal. That is, the star can be transferred from the d to the t, and therefore t and d^* are reciprocal, i.e., the translations of the lattice of the object and the spacings of the diffraction image are also reciprocal. That this is so can be quickly appreciated by imagining the direction of the light in Fig. 1 to be reversed. The diffraction from the reciprocal lattice is then the object lattice. Thus it is also true that

$$t_{uv} = (D_1\lambda_1)/d_{uv}^*. \tag{7}$$

(In the above discussion, the starred letters always refer to quantities in the diffraction image, i.e., to the reciprocal lattice.)

Furthermore, it is evident that, since the arrangement of spectra in the diffraction image is on points of a lattice, the forward diffraction of the diffraction image is also periodic. The same law as derived above holds, and the length of a period of the final image corresponds to that derived for a period of the diffraction image, but specifically

$$T = (D_2\lambda_2)/d^*. \tag{8}$$

The over-all magnification of the optical system is evidently

$$M = \frac{\text{length of a period in the image}}{\text{length of a period in the object}} = \frac{T}{t}. \tag{9}$$

Substituting in (9) from (8) and (7), the magnification turns out to be

$$M = (D_2/D_1) \times (\lambda_2/\lambda_1). \tag{10}$$

Now, in the usual optical system, the same light is used for both stages of diffraction, and (10) reduces to $M = D_2/D_1$, which is the usual ratio of image distance to object distance. But tremendous magnifications could be achieved if two very different wave-lengths

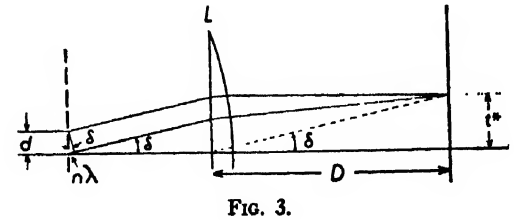


FIG. 3.

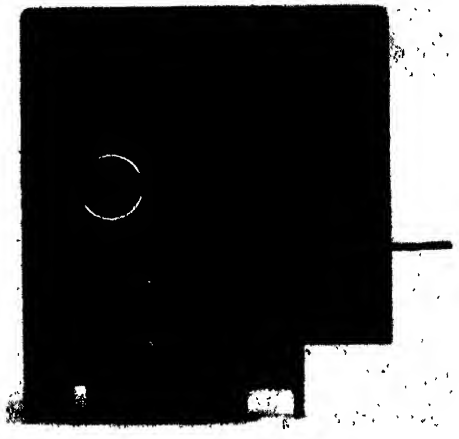


FIG. 4. Device for removing the Lorentz and polarization factors during the recording of de Jong-Bouman photographs. The large metal sheet with the annular opening acts as the layer-line screen. A sector rotates at high speed in front of the opening, propelled by an air jet. The interposing of the metal of the sector in the path of the reflections reduces the time during which each one records, by a factor which is proportional to the product of the Lorentz and polarization factors for the zero level.

were used for the two stages of diffraction. Specifically, if x-rays were used for the first stage, and visible light for the second stage, a factor of 10^4 could be realized apart from the further magnification due to choice of focal lengths of the lenses. Some specific values of the factors in the right of (10) which are used in the instrument described later are approximately

$D_2 = 2\times10^2$ cm
$D_1 = 5$ cm
$\lambda_2 = 5\times10^{-5}$ cm (green light)
$\lambda_1 = 7\times10^{-9}$ cm (molybdenum $K\alpha$-radiation).

With these values, (10) becomes

$$M = \frac{2\times10^2}{5} \times \frac{5\times10^{-5}}{7\times10^{-9}} = 3\times10^5 \text{ times.}$$

With this magnification, an atom 1A in diameter would appear as a circle 0.03 mm in diameter, which is large enough to see with a low power microscope. Such a scheme of magnification would have the advantage that *the resolving power of the system depends on the first wave-length, and consequently this enormous magnification is not empty.*

This system has two inherent disadvantages. In the first place, unless the first radiation of wave-length λ_1, arriving at the diffraction image, can be caused by some device to release the second radiation of wave-length λ_2 in appropriate phase, then the phase of each spectrum

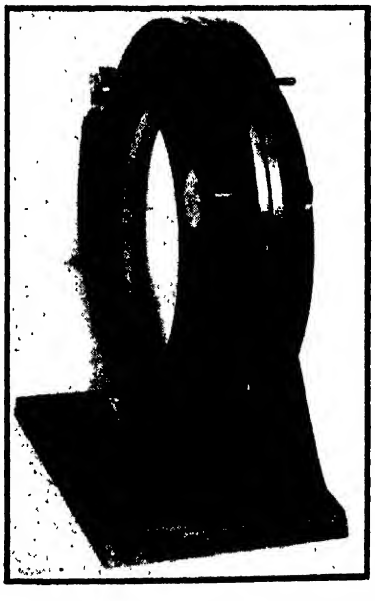


FIG. 5. Device for utilizing a print on film as a diffraction grating. The print is sandwiched between two thick glass flats, and the whole thickness is then rendered optically homogeneous by filling all space between film and flats with a liquid whose refractive index matches that of the film. The liquid is run into the system through the tube seen projecting at the upper right. The (forward) movable flat can be grasped by the transverse handles and removed after releasing the three screws. The clear aperture of the system is about 6 in.

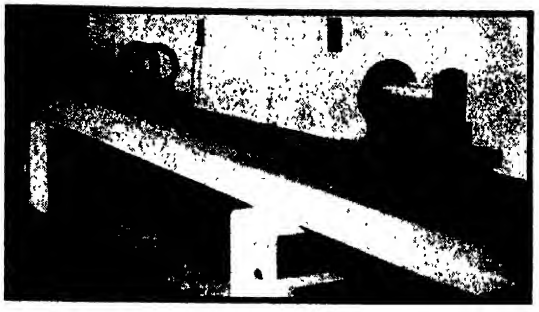


FIG. 6.

of the diffraction image is lost. At present there appears to be no obvious device for accomplishing this phased release of the second radiation. In the absence of such a device, the two-wave-length microscope requires that the phases be supplied. The mechanics of supplying the phases are described later.

Another inherent disadvantage of this system is that if a tri-periodic pattern, such as a crystal, is used as an object, the conditions for diffraction by a three-dimensional grating obtain. One aspect of such diffraction is that, in general, the spectra can only be obtained one at a time as the object is shifted in orientation to satisfy the Bragg condition for each plane one at a time. Thus, the diffraction from the diffraction image could not be made to interfere and build up a final image. Furthermore, if x-rays are to be used in the first stage, they cannot be focused by lenses. But fortunately, both of these difficulties can be avoided because there exist instruments for recording x-ray diffraction in a pattern which is arranged precisely as if the pattern *had* been focused by lenses and as if the several diffraction spectra *had* been produced at once. This is accomplished by the de Jong-Bouman apparatus[3] or by the precession apparatus.[4] The law of magnification for these instruments is precisely (5), where D_1 is the crystal-to-film distance, and λ_1 is the x-ray wavelength.

Thus, in effect, the system *can* be used to form a magnified image of the patterns of crystals provided the appropriate phases of the diffraction by the diffraction image can be supplied. To use the system, it is only necessary to substitute at the plane of the first diffraction image, the positive photograph of the x-ray diffraction pattern (or its equivalent) arranged in reciprocal lattice form by one of the devices mentioned in the last paragraph, and then illuminate this pattern with visible light. In this way, in effect, visible light is substituted for x-rays in the right part of Fig. 1. The philosophy of generalized microscopy is then this: Record (for example, by photography) the diffraction image of the first stage of image formation using radia-

[3] See reference 2, pp. 331–346.
[4] M. J. Buerger, "The photography of the reciprocal lattice," Am. Soc. X-Ray and Electron Diffraction, Monograph No. 1, pp. 1–37 (1944).

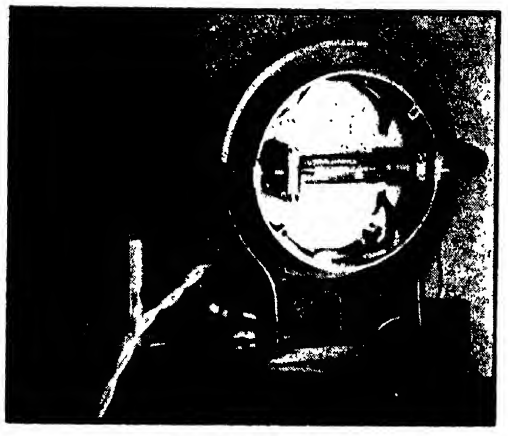


FIG. 7(a). Water-cooled mercury arc source.

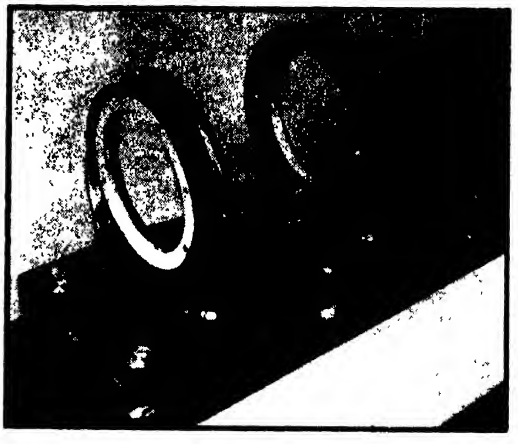


FIG. 7(c). Lens assembly.

FIG. 7(b). Mercury arc source focused on a pinhole of the pinhole assembly.

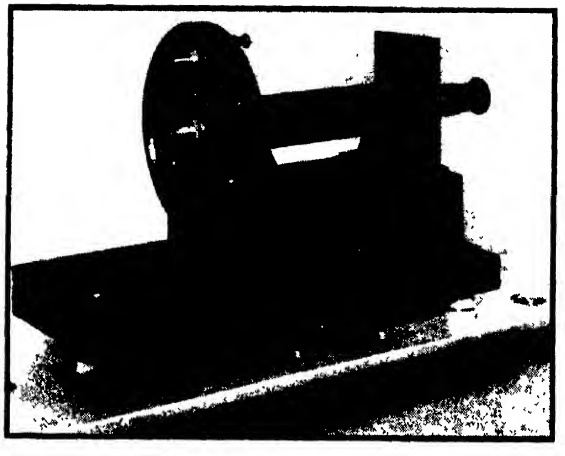


FIG. 7(d). Microscope assembly.

tion of short wave-length; then, placing the positive photograph (or equivalent) of this diffraction image at the correct position of the diffraction image, illuminate it with radiation of longer wave-length. After providing the right phases, the image now obtained is the same as if the original radiation had been used throughout, except that the magnification is scaled up by a factor λ_2/λ_1.

The use of positive photographs of x-ray diffraction arranged in reciprocal lattice form has been practiced by the writer. It has several disadvantages. In the first place, the x-ray diffraction cameras distort the intensities of the spectra by multiplying each by the Lorentz and polarization factors.[5] Fortunately this distortion can be automatically removed during recording by a rapidly rotating sector,[5] shown in Fig. 4. Second, the photographic film surfaces are too irregular for diffraction work. This can be offset by immersing the film in a liquid of its own refractive index, and sandwiched between two optical flats utilizing apparatus such as shown in Fig. 5. This has also been successfully practiced.

But, in spite of its inelegance, a more satisfactory way, from a mechanical point of view, is to determine the intensities of the x-ray diffraction spectra by the Dawton technique[6] (correcting for Lorentz and polarization factors if this has not already been done automatically during recording) and make a mechanical equivalent of the positive diffraction pattern consisting of a metal plate with a hole at each spectral position whose area is proportional to the magnitude of the amplitude $|F_{hkl}|$ of the spectrum. A large number of unit cells of the structure can be accurately imaged in the focused diffraction of such a grating if the largest hole does not exceed about 50 wire gauge (0.070 in. diameter). This equivalent of the first diffraction image has mechanical qualities which lend themselves well to the phase-shifting arrangements to be described.

A view of the assembled optical units of the apparatus which represents the second diffraction stage of Fig. 1

[5] M. J. Buerger, "The correction of x-ray diffraction intensities for Lorentz and polarization factors," Proc. Nat. Acad. Sci. 26, 637–642 (1940).

[6] M. J. Buerger, "The photography of interatomic distance vectors and of crystal patterns," Proc. Nat. Acad. Sci. 25, 383–388 (1939).

is shown in Fig. 6. This comprises, from left to right, a source of light, a pinhole, two lenses, and a microscope. The light source illuminates the pinhole, which acts as a point source of light. The diverging radiation from this source is collimated by the first lens and strikes the reciprocal-lattice diffraction grating as a plane wave. The diffraction from the grating is focused by the second lens so that the final image occurs near the right side of the assembly. The device at the extreme right is a low power microscope which magnifies the image so that its details are easily visible to the eye.

The units of the optical system are shown in Fig. 7. They are mounted on individual beds. Each unit and bed is fitted with a millimeter scale and vernier attachment so that the unit, if removed for any purpose, can be accurately returned to its correct location.

The optical path of the system is normally shielded from the disturbing influences of stray light by a tubular cover, seen in Fig. 8. The optical system is also fitted with an auxiliary mirror which can be raised into the cone of rays converging from the second lens. This throws the diffraction image back to a point near the grating, where it can be viewed by the operator from this region, Fig. 9. This arrangement permits the operator to make any desired adjustments on attachments to the grating, such as the phase shifters to be described beyond, while watching the results.

The source of light is a water-cooled mercury arc, specifically the General Electric *A-H6*. The mercury green line (5461A) is isolated from the arc spectrum. Shorter wave-lengths are eliminated by a Wratten filter No. 12, whose cut-off is nearly complete below about 4900A. This is placed permanently behind the lens of the microscope objective. Longer wave-lengths, particularly the orange component near 5800A, are eliminated by taking the photographs on orthochromatic film.

The mercury arc illuminates a very small pinhole. The limitations on the size of this pinhole can be appreciated if one notes that, with the magnification mentioned earlier, an atom one angstrom in diameter would appear as a disk about 0.03 mm in diameter. This is about 0.001 in., and the upper limit of the diameter of the pinhole must not substantially exceed this, and preferably it must be considerably less in order to resolve detail in the individual atoms of the image. In the apparatus shown in Fig. 6 there is a turret device, Fig. 7(b), which permits one to swing pinholes of four different sizes into position. The two largest pinholes are 0.0015 and 0.0007 in. in diameter, respectively. These were prepared by drilling holes mechanically in thin pieces of brass sheet. Their principal use is to examine images for atomic position, where atomic detail is not required. The turret device also bears two very much smaller pinholes of graded sizes which were found to occur fortuitously in thin aluminum foil. These are necessary when detailed photographs, such as shown in Fig. 13, are to be taken.

The two lenses are plano-convex and are substantially duplicates, with diameters of 6 in. and focal lengths of 6 ft. They are very perfectly corrected for spherical abberation but, of course, require no correction for chromatic abberation since they are used with monochromatic light.

The microscope at the right end of Fig. 6 has two standard parfocal microscope objectives of 16 and 32 mm focal lengths, respectively, mounted in a special turret so that they can be quickly interchanged, Fig. 7(d). The microscope image can be either examined visually through an eyepiece, or a small film holder can be substituted for the eyepiece, Fig. 6, and the image photographed. As mentioned above, orthochromatic film is used for this purpose in order to eliminate the orange mercury line.

It has been pointed out[1] that any phase shift whatever can be imposed on the rays scattered by a point of the reciprocal grating by interposing in the path of the rays going through the point a tilted flake of mica which has been ascertained to have a truly uniform thickness. This was experimentally proven by the writer to be feasible as early as 1941. For this purpose, a thin cleavage flake having as large an area as possible is separated from an excellent crystal of mica. The thickness uniformity is easily judged by a very simple interferometer method: The flake is merely held close to a strong broad source of monochromatic light, such as a General Electric "Labarc." Interference fringes then appear to contour the mica flake, and these are without discontinuities in fields where the cleavage flake has a uniform thickness. It is a simple matter by

FIG. 8.

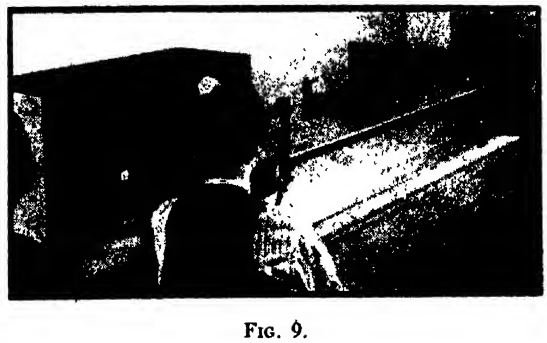


FIG. 9.

FIG. 10.

such observations to map out the regions of the flake which are uniform in thickness. A large uniform area is then selected for the purpose of cutting out the phase shifter units described below.

In order to calibrate the mica sheet so prepared, a small sample of the sheet is cut out and mounted on a rotation device which is placed so that the mica is directly in front of one of two equal circular holes of a two-hole "grating," shown in Fig. 10. The diffraction pattern of the pair of holes can be examined by placing the "grating" in the optical apparatus described above. Without the mica in the system, the diffraction pattern consists of a series of short linear maxima whose long directions are normal to the separation direction of the two holes. When the mica sample is placed in front of one of the holes, these maxima are shifted, in general, in accordance with the phase shift induced by the additional path length through the mica. This path length, and its attendant phase shift, is a function of the angle which the normal to the mica sheet makes with the rays passing through the hole. By measuring the drift of the fringes as a function of the mica angle, the phase shift produced by a particular mica flake can be determined as a function of this angle. For example, the setting of the mica which causes a shift of maxima so that one maximum comes to be located where a minimum occurred with no mica in place, corresponds to a phase shift of π. Proceeding in this way, the angle which the mica should make with the optical axis of the system to cause a phase shift of any required amount can be determined. The theory and practice of constructing phase shifters will be discussed in more detail in another paper.

A practical way of placing the mica sheets at the correct angles in front of the various holes of the grating is illustrated in Fig. 11. This particular example shows phase shifters in place in front of the grating used for producing the image of marcasite, FeS_2, described later and shown in Fig. 13. The individual mica pieces are mounted on small brass blocks in such a way that the mica makes the required angle with the rays through the grating holes. The blocks are placed on shelves along each row of one set of rows in the reciprocal lattice.

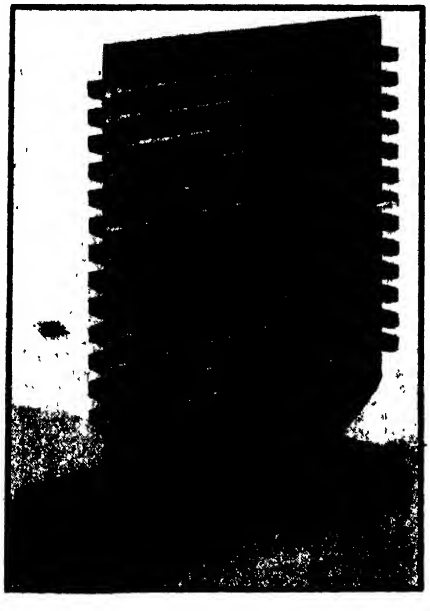


FIG. 11.

FIG. 12. Left to right: (a) Assembly for producing phase shift of π for centrosymmetrical images. (b) Assembly for producing any phase shift required for non-centrosymmetrical images. (c) and (d) The assembly shown in (b) taken apart. The piece in the right is the mica-bearing member whose orientation can be varied within the metal block seen in (c).

It is convenient to have two kinds of phase-shifter blocks. The rather simplified block shown in place in Fig. 11 is desirable when the pattern of the crystal structure projection to be imaged is centrosymmetrical. In such cases the phases can have only the values 0 and π, so only two discrete values of phase must be impressed on the scattering from the holes of the grating. The phases of the waves scattered by those holes whose scattering phase is to be set at zero need not be regulated, for the phase of the plane wave impinging on the grating may be taken as zero. But the phases of the waves scattered by holes whose scattering phase is to be set at π must be controlled by placing in front of each such hole a block bearing a piece of mica set at the correct angle to produce the phase shift of π. The design of this simple phase shifter is shown in Fig. 12. To image an average crystal structure, about 100 phase shifters are ordinarily needed. This requires that the original piece of mica should be of sufficiently large area to permit cutting 100 duplicate pieces from it. Each such fragment is cut so that the rotation of the

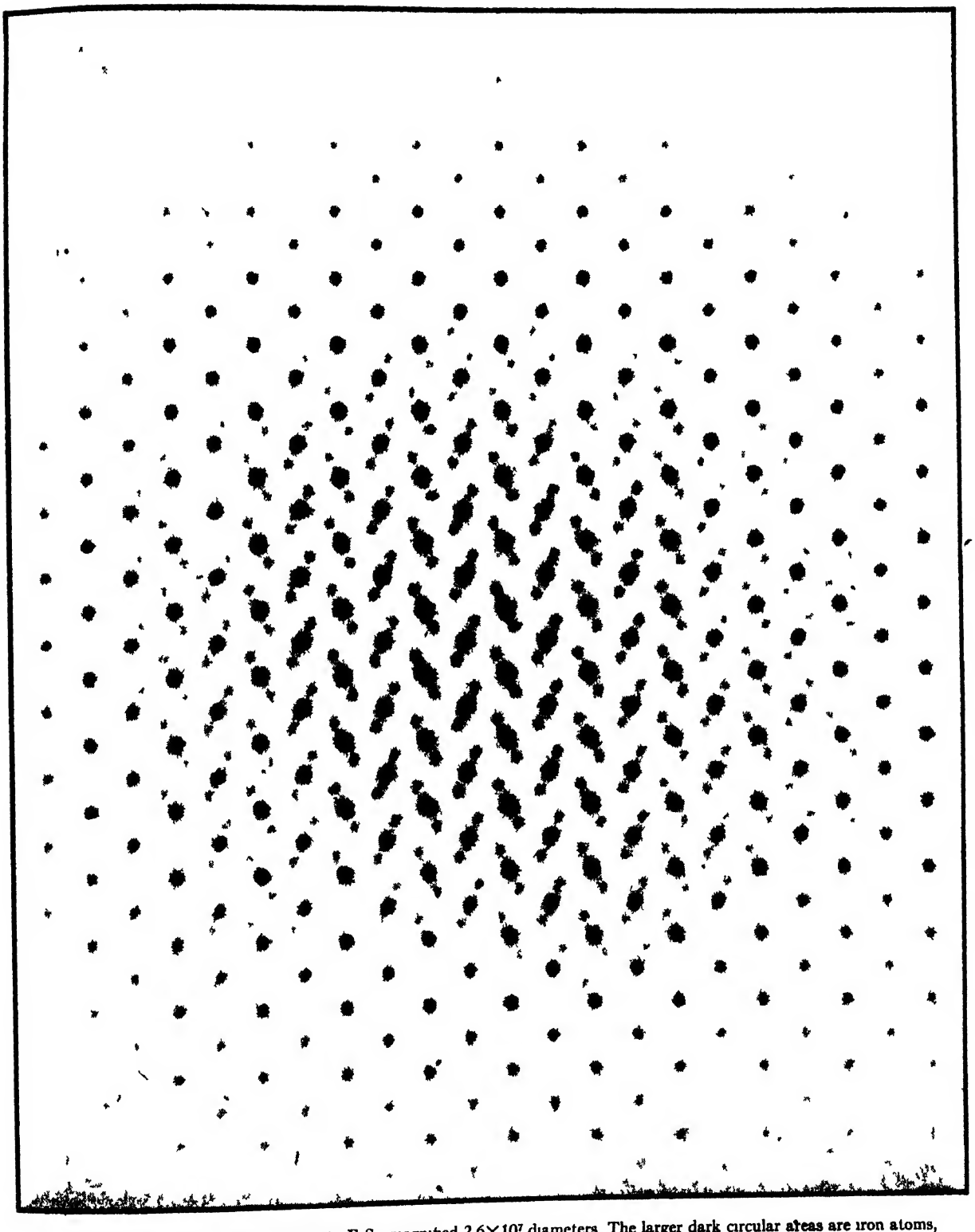


FIG 13 Image of the atoms in *marcasite*, FeS_2, magnified 2.6×10^7 diameters The larger dark circular areas are iron atoms, with 26 electrons each while the fainter darkened areas are sulfur atoms with 16 electrons each

mica is carried out about its optical direction Y, which minimizes the double refraction of the mica.

For imaging projections of crystal structures without centrosymmetry, each spectrum has, in general, its own different phase. To impress all these different phases that the phase-shifter block must have, the mica must be mounted so that its angle with the optical axis of the system can be varied. Figure 12 shows the design of such phase shifters which has been found to be practical. Between 200 and 400 of such units are needed to properly image the average crystal structure projection.

Figure 11 showed the phase shifters in place on the grating appropriate for imaging the crystal structure of marcasite, FeS_2 as seen along the c axis, and Fig. 13 shows the final image produced by the two-wave-length microscope system, utilizing this grating. The iron atoms, with 26 electrons each, are imaged as the heavier spots, while the sulfur atoms, with only 16 electrons each, are imaged as the lighter spots. The photograph is an accurate electron density map of this crystal structure.

To produce this image, the 150 individual $hk0$ spectra within recording range of the 30° precession photograph made with Mo$K\alpha$-radiation were utilized. A few of these spectra had amplitudes of zero. Of the entire 150 possible spectra, 32 required phase shifts of π, and Fig. 11 shows the 32 corresponding phase shifters. The test of the success of the phase shifting is the reproduction of the sulfur atoms, which are imaged as the less dense spots. To image the iron atom positions in this particular pattern requires the use of positive phases only, but to image the sulfur atom positions properly requires that the correct combination of negative and positive phases has been impressed on the grating.

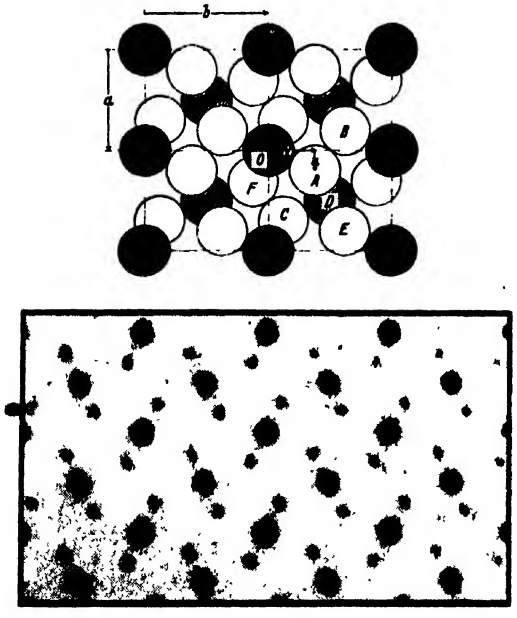


FIG. 14. Below: Image of the atoms in *marcasite*, FeS_2. Above: Diagrammatic representation of the crystal structure of *marcasite*.

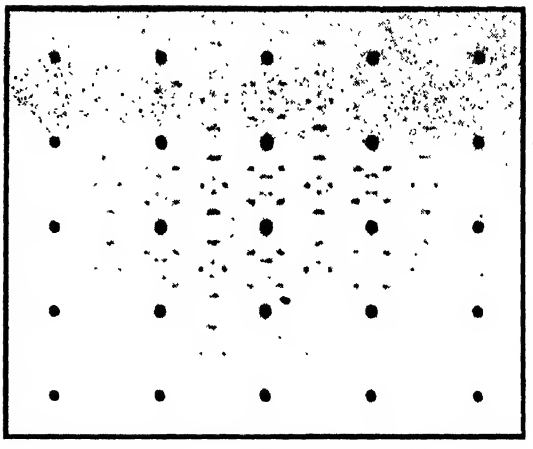


FIG. 15. Photographic Patterson synthesis, $P(xy)$ for *berthierite*, $FeSb_2S_4$.

Figure 14 shows part of Fig. 13 compared with a diagrammatic representation of the crystal structure of marcasite taken from an earlier paper.[7] In the diagrammatic representation, the shadings of the circles do not represent electron density, and their sizes merely illustrate the packing radii of the atoms, which are not related in a simple way to the electron densities in the atoms. With this proviso, the photographic reproduction of the atoms is in striking agreement with the published crystal structure of marcasite.

A word of caution is required. The development just described *does not* mean that now any crystal structure can be photographed. It *does* mean that this can be done provided that the phases which must be set with the mica phase shifters is known. These are the phases of the F's of x-ray diffraction, and they cannot be determined experimentally, although the magnitudes of the F's can. To know the phases requires, in effect, that the locations of the atoms are known. But when the phases are not known, which is the state of affairs as a new crystal structure is first examined, the apparatus described above automatically gives the Patterson synthesis[8,9] of the crystal, a fact which was noted[6] in 1939. An example of such a synthesis is shown in Fig. 15. The Patterson synthesis is related to the true image of the crystal structure in the following way: If one views the crystal structure from a fixed point, he sees a true image of the crystal structure. But if one occupies each of the n atoms of the structure in turn, viewing the crystal structure from each one, and then superposes all n parallel images, the composite image is the Patterson synthesis. Theoretical progress has recently

[7] M. J. Buerger, "Interatomic distances in marcasite and notes on the bonding in crystals of löllingite, arsenopyrite, and marcasite types," Zeits. f. Krist. (A)97, 504–513 (1937).

[8] A. L. Patterson, "A Fourier series method for the determination of the components of interatomic distances in crystals," Phys. Rev. 46, 372–376 (1934).

[9] A. L. Patterson, "A direct method for the determination of the components of interatomic distances in crystals," Zeits. f. Krist. (A), 517–542 (1935).

been made in the direction of eliminating from the Patterson synthesis all crystal structure images but one.[10] When this theory and practice is perfected, it will become possible to set the phases for producing the true image by utilizing the first Patterson image made without phases.[11]

The writer is indebted to the National Research Council Committee on Scientific Aids to Learning for a grant in 1939 to cover the construction of the instrument described in this paper. He is also personally grateful to Drs. Vannevar Bush and Warren Mead for their encouragement in the development. The undertaking was greatly aided by the assistance of Mr. John Tyler, who superintended the engineering of the main optical system, turning ideas into realities. The gratings, phase shifters, and pinholes are the expert handiwork of Mr. John Solo, and the writer is also indebted to a graduate student, Mr. Jay W. Lathrop, for examining and calibrating suitable mica for the phase shifters used to make Figs. 13 and 14.

[10] M. J. Buerger, "Vector sets," Acta. Crystal. 3, 87–97 (1950).
[11] M. J. Buerger, "Some new functions of interest in x-ray crystallography," Proc. Nat. Acad. Sci. 36, 376–382 (1950).

A Model of Hurricane Formation*

HERBERT RIEHL
University of Chicago, Chicago, Illinois
(Received April 10, 1950)

At first, the description of the energy cycle of the mature tropical storm is amplified on the basis of recent upper air observations. Air particles passing through such a storm at first undergo isothermal expansion as they move toward a center. Then they ascend with release of condensation. At high levels they move outward and mix with the environment giving off heat to the surrounding colder air. Several requisites for maintenance of the observed temperature field are stated.

After a discussion of previous theories of hurricane formation the proposed model is described. The initial intensification of the wind field is brought about by mass divergence at high levels that imposes a pressure reduction on the surface layers. This divergence is the result of interaction between the large-scale disturbances of the upper air inside and outside the tropics. A solenoidal circulation is initiated that acts in the kinetic energy producing sense. But this circulation contains an internal mechanism for its own destruction and is maintained only under certain special conditions which are stated.

SINCE the early 1940's, numerous observations of pressure, temperature and wind have been taken in the atmosphere in the tropics at heights up to 20 km. These make possible a re-examination of the views held concerning the formation of tropical storms. Such a re-examination is the object of this report. New observational data are not offered. For these, the reader is referred to several of the references listed.[1–3]

THERMAL STRUCTURE OF FULLY DEVELOPED STORM

It is of advantage to begin with comments on the hurricane circulation after it has grown to full intensity. In some ways, tropical storms are those disturbances within the atmosphere whose maintenance is easiest to understand at present. Figure 1 is a representative sketch of pressure, temperature and circulation in a vertical section across a well-developed hurricane or typhoon outside the "eye." This figure shows a simple heat engine. Near the surface, air moves from high toward low pressure, converges and ascends. During the ascent, latent heat of condensation is liberated. In consequence the air near the center is warmer than its surroundings through a deep layer and the vertical spacing between the surfaces of constant pressure is larger. The pressure gradient force, directed toward the center, decreases upward and eventually reverses. Air particles moving upward near the center will be accelerated outward throughout the layer of decreasing pressure gradient, if their angular momentum is conserved.[4]

Figure 2 shows the vertical structure of the air surrounding tropical storms and a characteristic curve of the structure inside the rain area on a thermodynamic chart. Curve (b) is the same as that obtained by raising an air particle dry-adiabatically from sea surface to condensation level, and then moist-adiabatically to the 200-mb level. *It follows that the air mass encountered in the rain area below 200 mb consists entirely of particles that have ascended from the lowest levels.* Entrainment of mass from the surroundings does not take place, since such entrainment would produce upper temperatures considerably colder than those of curve (b). The air that

* A report on research conducted under contract between the ONR and the University of Chicago.
[1] H. Riehl, "On the formation of West Atlantic hurricanes," Dept. Meteorolog. Univ. Chicago, Misc. Rep., No. 24, (1948) 1–64.
[2] H. Riehl, "On the formation of typhoons," J. Meteorolog. 5, (1948) 247–264.
[3] H. Riehl and N. M. Burgner, "Further studies of movement and formation of hurricanes," Bull. Am. Meteorolog. Soc. 31 (No. 7), (1950).
[4] C. S. Durst and R. C. Sutcliffe, "The importance of vertical motion in the development of tropical revolving storms," Quart. J. Roy. Meteorolog. Soc. 64, (1938) 75–91.

fills the body of a hurricane originally is spread out in a thin sheet along the surface of the earth. This sheet—in which potential temperature and moisture content are nearly constant[5] has a depth of 50–60 mb. If a storm has a radius of 100 km, and if it extends to the 200-mb level, the surface air must be drained off from an area with a diameter of about 800 km in order to produce the first formation of the air mass observed inside the storm. This is an appreciable value.

One aspect of the temperature field has not received the attention it deserves. Byers[6] pointed out that the air spiralling inward toward a center experiences considerable expansion because of the pressure reduction which may have values from about 30 mb up to 100 mb. The temperature therefore should drop at a rate of nearly 1°C per 10 mb pressure reduction on account of adiabatic expansion. *Such decreases never occur. On the contrary, surface temperatures are remarkably constant throughout the body of tropical storms. The air moving inward expands at constant temperature,* and the potential temperature must increase from outskirts toward center. Rapid transfer of heat from ocean to atmosphere makes the isothermal expansion possible. In view of the highly agitated state of the ocean in hurricanes the occurrence of such rapid transfer is easily understood. Computations in a number of representative cases, mostly taken from data published by Deppermann[7], have verified the potential temperature increase. They have shown, moreover, that the moisture content of the air

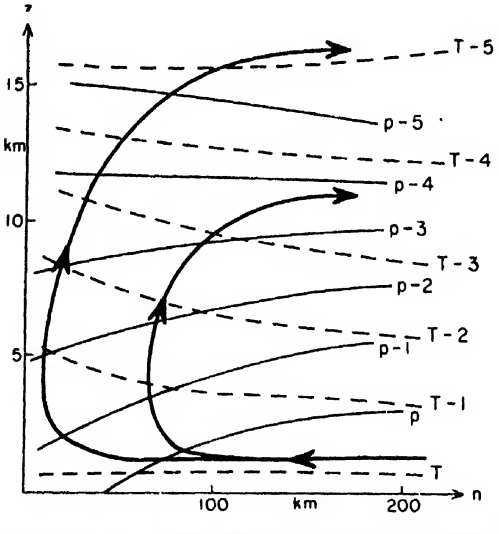


Fig. 1. Schematic outline of vertical circulation and distribution of temperature (T) and pressure (p) in rain area of a mature hurricane. The "z" axis points upward and the "n" axis outward across the storm from the eye boundary.

[5] Bunker, Haurwitz, Malkus, and Stommel, "Vertical distribution of temperature and humidity over the Caribbean Sea," Pap. Phys. Ocean. and Meteorolog. Mass. Inst. of Tech. and Woods Hole Ocean. Inst. 11, No. 1, (1949).

[6] H. R. Byers, *General Meteorology* (McGraw-Hill Book Co., Inc., New York, 1944).

[7] C. E. Depperman, *Some characteristics of Philippine typhoons* (Bureau of Printing, Manila, 1939).

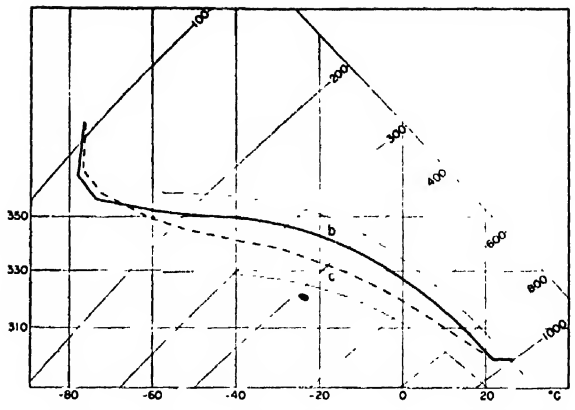


Fig. 2. Tephigram of (a) mean tropical atmosphere after Schacht (1946) and (b) sounding in rain area. Vertical lines are isotherms (°C), lines slanting from upper right to lower left isobars (mb) and horizontal lines are lines of equal potential temperature (dry-adiabats °A). Two moist adiabatic lines are entered on both sides of the soundings.

also becomes larger by 1 to 2 grams of water vapor per kg of air. This increase is also readily understood.

Figure 3 demonstrates the importance of this local energy source in tropical storms. Curve (b) represents the ascent of the surface air without local heat intake as in Fig. 2. Curve (a) shows an ascent from 960 mb under the assumption that the moisture content has risen by 1.5 g/kg and that the air has expanded isothermally to 960 mb. The difference between the two ascent curves is large. Comparison between curves (a) and (b) in Fig. 2 and curve (a) in Fig. 3 shows *that the area enclosed on the thermodynamic chart between the normal surroundings and the interior of a storm doubles in size if we allow for the local heat source.*

We must now turn to the problem of maintenance of the temperature field sketched in Fig. 1. The streamlines shown in this figure indicate outward transport of warm air aloft which should raise temperatures in the outskirts and destroy the solenoid field. If we make use of the local heat addition just mentioned, it is possible to obtain a temperature field that is partly steady. Let the air ascending near the outskirts of a cyclone follow curve (b) of Fig. 3, and the air nearest the center curve (a). Values intermediate between these extremes will be found in the middle region. If the air following curve (a) ascends the largest distance and the air following curve (b) the least, as sketched in Fig. 1, then the horizontal temperature gradient inside the storm will be maintained in the low and middle troposphere.

Outside the boundary of the storm, the outward transport of very warm air should be observed within a radius of 500 miles and more soon after establishment of a hurricane because of the large mass transport involved. Actually, however, the extremely warm air is not noticed at such a distance from the center. A slow and gradual warming of the surroundings takes place and very high temperatures are never reached. This

permits only one conclusion, namely that the air flowing out at high levels mixes with the surrounding air masses, that have not passed through the circulation, and gives up heat to them. At the same time the mixture is carried away from the storm area by the currents aloft and replaced by cooler air from other portions of the globe. In this way the entire temperature field relative to the storm is maintained. Two things must be noted. The mixing must take place outside the core since, as shown during the discussion of Fig. 2, the air of the environment does not enter the center aloft. It is also necessary to have a mechanism that is able to carry the mixture far away from the storm area and permit other air to approach.

Three phases of the energy cycle of the air passing through the cyclonic circulation have now been described. They are: (1) isothermal expansion with moisture addition; (2) dry-adiabatic expansion to the condensation level, followed by moist-adiabatic expansion; (3) mixing with cooler air aloft, a process that takes place at nearly constant pressure.

Figure 4 illustrates these three steps. For the construction of this figure a sample of air has been chosen that ascends to 200 mb and mixes with the surroundings at that level. The reduction of temperature from −47°C to −52°C at 200 mb, as indicated, is plausible, if we consider that the surroundings normally have a temperature of −55°C which is observed to rise to about −52°C in the situation under discussion. It is readily apparent from Fig. 4 that only a small percentage of the heat released during expansion of the ascending air from 1000 mb to 200 mb is available for conversion into kinetic energy. Maximum conversion is obtained if the air descends dry-adiabatically (following the dashed horizontal line) to the level at which its temperature becomes equal to that of the interior of the storm. The upper area (a) in Fig. 4 is only 12 percent of the total area (a+b) enclosed by the hurricane sounding and the normal surroundings. This represents the maximum efficiency of the circulation which, however, is never realized. As stated before, temperatures comparable to those inside are not observed outside. Therefore only a

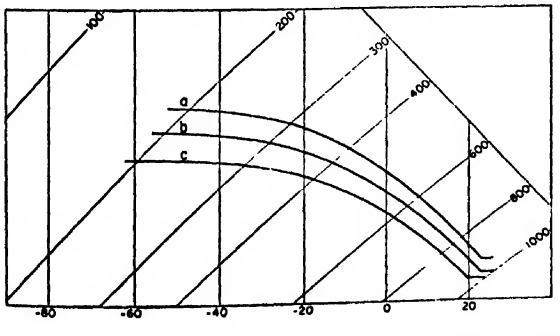


FIG. 3. Tephigram showing (a) ascent of air that has expanded isothermally to 960 mb at the sea surface with pickup of 1.5 g/kg of moisture; (b) same as sounding (b) of Fig. 2; (c) ascent of air with slightly polar characteristics (24°C, 16 g/kg).

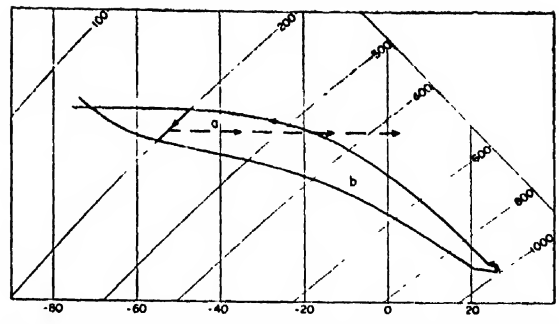


FIG. 4. Tephigram showing isothermal expansion of air at sea surface to 960 mb, dry-adiabatic ascent to condensation level, then moist-adiabatic ascent to 200 mb and horizontal mixing at 200 mb. Dashed horizontal line indicates descent of mixture outside the circulation. Lower curve is same as curve (a) of Fig. 2. The area (a+b) represents total heat gained by hurricane circulation, and area (a) represents maximum amount of energy available for conversion into kinetic energy.

fraction of the energy represented by area (a) actually is used to drive the storm. The vast bulk of the heat released is spread over the globe.

Figure 4 shows clearly that a tropical storm and its surroundings cannot be considered as a closed system. It is a matter of weeks or more until the air that has risen in the storm can return to the surface under the influence of radiation. New air constantly enters the cyclone. In the course of the life of one storm, the air initially located over a large portion of a tropical ocean may pass through its circulation system.

Using the preceding discussion as a base, we are now in a position to examine critically several theories of hurricane formation that have been put forward.

PREVIOUS THEORIES OF HURRICANE FORMATION

Convectional Hypothesis

This is the oldest hypothesis and briefly states the following: an unusually large number of heavy rain squalls and thunderstorms develops for some reason over tropical ocean. These rain clouds then grow together and a cyclonic circulation develops at the ground due to convergence. Because of the heavy rainfall energy is available for an increase of the circulation (curve (b) of Fig. 2).

Most writers have discarded this hypothesis as insufficient for a number of years. But it still recurs in some textbooks. Here we shall mention only one objection: a mechanism for organized mass removal aloft is not specified, so that the area of heavy showers and its immediate surroundings will act almost as a closed system. After ascending the air will descend dry-adiabatically in the environment (following the horizontal lines of Fig. 2) and the environment will soon be warmer than the cloudy area. Any incipient circulation must dissipate. Actually, most heavy rainfalls of the tropics occur when the wind field is very weak or only slightly cyclonic. A sequence of events then takes place

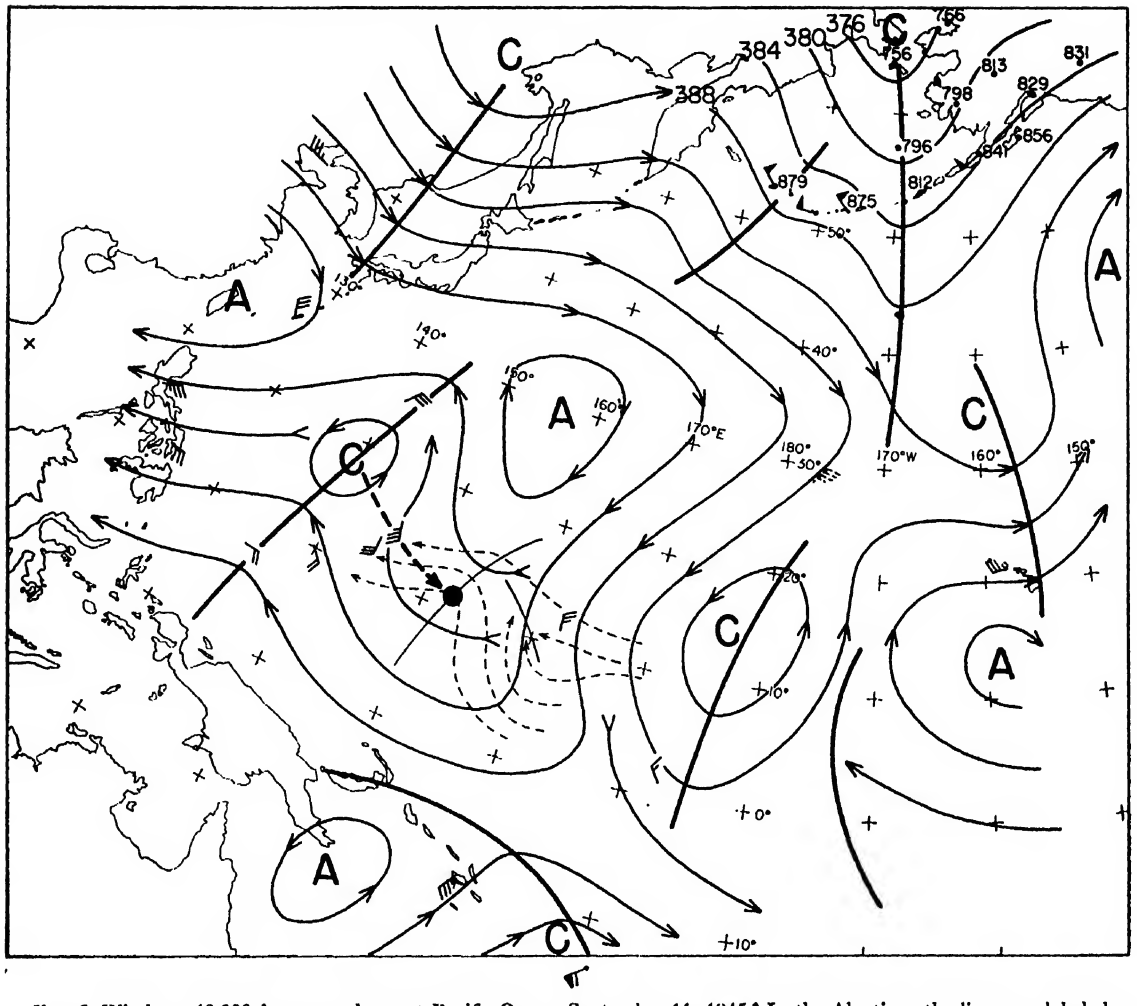


FIG. 5. Winds at 40,000 feet over the west Pacific Ocean, September 11, 1945.[2] In the Aleutians the lines are labeled as contours of the 200 mb surface (hundreds of feet). "*C*" stands for cyclonic circulation, "*A*" for anticyclonic circulation, and heavy solid lines indicate wave troughs and shear lines. In the wind reports a long barb denotes 10 m.p.h., a short barb five m.p.h. and a heavy triangular barb 50 m.p.h. Near the Marianas the 10,000-foot flow is shown by light dashed lines. Heavy dashed line denotes motion of incipient typhoon (heavy dot) relative to upper disturbance in which it had started during several preceding days.

which is fairly similar to that just outlined. Energy must be drawn into these circulations from an outside source in order to maintain them.

Frontal Hypothesis

About 1920, the Norwegian meteorologists formulated a theory of formation of shearing waves at density discontinuities (fronts). It has been since suggested that the zone of convergence, where the trades from northern and southern hemisphere meet, also is a boundary with a small density discontinuity along which shearing instability can develop. Actually, many storms, though by no means all of them, originate near this boundary. It is not possible, however, to accept the frontal hypothesis as an explanation.

TABLE I.

P	1	2	3	4
surface	26.0	26.0	26.0	24.0
900	20.0	20.5	22.5	18.0
700	8.0	11.0	13.5	8.0
500	−6.0	−3.0	0.5	−6.5
300	−33.0	−28.5	−23.5	−34.0
200	−55.0	−53.0	−47.0	−59.0

As in the case of the convection theory, a mechanism for outflow is not provided. There is another serious objection. According to the Norwegian concept, the denser air mass settles underneath the lighter mass, and this release of potential energy can lead to an increase of the kinetic energy. After some time (24–36 hours) the cyclone core at the surface consists entirely of cold air. Most of this cold air sinks in case of cyclones outside the tropics. In fundamental contrast, all air that enters a tropical cyclone near the surface ascends in the core. If a tropical cyclone core becomes filled with air that is even

slightly denser—and usually somewhat drier—than the normal tropical atmosphere, conditions become extremely unfavorable for the growth of a storm. Curve (c) of Fig. 3 illustrates this difficulty. This curve gives the ascent path of air that is very little cooler and drier ($T = 24°C$, specific humidity = 16 g/kg) than the air used for the construction of curve (b) ($T = 26°C$, specific humidity = 18 g/kg). A large temperature difference aloft results. If we compare curve (c) of Fig. 3 and curve (a) of Fig. 2 we see that no positive area is left on the thermodynamic chart. In fact, there is now a slight negative area. For comparison, the data of Figs. 2 and 3 are presented below in tabular form. Temperatures at certain isobaric surfaces are given for: (1) the average tropical atmosphere in the Caribbean; (2) the ascent of the surface air of average properties (26°C, 18 g/kg); (3) the ascent of the same air after isothermal motion to 960 mb with pickup of 1.5 g/kg of moisture; (4) the ascent of air with very slight polar characteristics (24°C, 16 g/kg).

Clearly, the entry of air with slightly polar characteristics must stop any incipient storm. It is a good forecasting rule that even intense hurricanes will weaken if they are affected by polar air while over a tropical ocean.

Hypothesis of "Dynamic Instability"

The foregoing theories are open to still another objection. Both take low level flow convergence as a starting point. Such convergence, however, must lead to pressure rises and not falls at the center, thus killing the beginning circulation. This is a well-known difficulty in cyclone theories. It is necessary to assume that the upper outflow somehow becomes sufficiently large to overbalance the surface inflow. To date, it has not been demonstrated how this is to be accomplished.

It is much more natural to suppose that at first a dynamic pressure fall occurs at the ground, produced by divergence at higher levels. Thereafter low level convergence sets in induced by the imposed pressure field. Sawyer[8] presented a hurricane theory based on this line of attack. He followed the work of Solberg[9] who showed that a symmetrical ring of air can become "dynamically unstable" under certain conditions. This can happen when the rotation of the air in space about the vertical axis (normal to the earth's surface at the place in question) is in the sense opposite to that of the rotation of the earth itself. Consider a ring of particles moving from west toward east along a latitude circle. If Solberg's criterion is satisfied, particles that are accelerated northward or southward will not tend to return to their starting latitude but move farther away from it. Under such circumstances dynamic pressure reductions can be produced.

Although Sawyer's attempt represents considerable progress, it is not conclusive for two main reasons. It is not indicated how the requisite sense of rotation is produced, nor is it shown how the unstable state is maintained. Even if "dynamic instability" exists in the region above a forming storm, the air evicted by this mechanism will sink in the surroundings. We then face the same reversal of the solenoid field which had made it impossible to accept the convectional hypothesis. No treatment exists as yet concerning the production of "dynamic instability." In middle latitudes, there frequently is a narrow zone in the high troposphere south of the maximum westerlies in which Solberg's criterion is satisfied.[10] Within the writer's experience, such a zone cannot be found in low latitudes, at least north of 15°N, in spite of the small value of the rotation of the earth itself (Coriolis parameter). Since data are scarce and scattered in the tropics, it is not easy to prove this assertion. Even at latitude 20°, however, a wind shear of 4 m sec.$^{-1}$/deg. lat., or a radius of flow curvature of 2° latitude at a windspeed of 10 m sec.$^{-1}$, are necessary in the limiting case when the rotation of the air particles relative to the earth is just equal in magnitude to the rotation of the earth itself. No plausible mechanism has been offered for the production of such large values over appreciable areas in the tropics prior to cyclone formation.

In summary, the theories just examined have one principal common drawback: they all consider only internal characteristics of the tropical atmosphere in the area of an initial disturbance. After some intensification this disturbance apparently is thought to be self-propagating. As we have seen, this is impossible. *At all times there must exist a connection between a hurricane and the surrounding atmosphere aloft over wide distances.* Sawyer's attempt is attractive in that it seeks to develop an initial field of upper divergence rather than a field of low level convergence. In the following, we shall retain this approach. But it is more plausible, and suggested by the observations, to specify an external rather than an

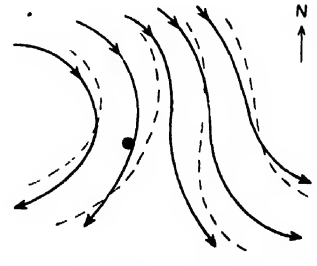


FIG. 6. Model of streamlines and contours (dashed lines) at 200 mb during deepening of a tropical disturbance (heavy dot) in northern hemisphere.

[8] J. S. Sawyer, Notes on the theory of tropical cyclones, Quart. J. Roy. Meteorolog. Soc. 73, (1947) 101–126.

[9] H. Solberg, "Le mouvement d'inertie de l'atmosphère stable et son rôle dans la théorie des cyclones," Procès-verbaux de l'assoc. de meteorolog. Un. geod. geophys. Int., Edinbourg, (1936) 66–82.

[10] University of Chicago, Department of Meteorology, On the general circulation of the atmosphere in middle latitudes, Bull. Am. Meteorolog. Soc. 28, (1947) 255–280.

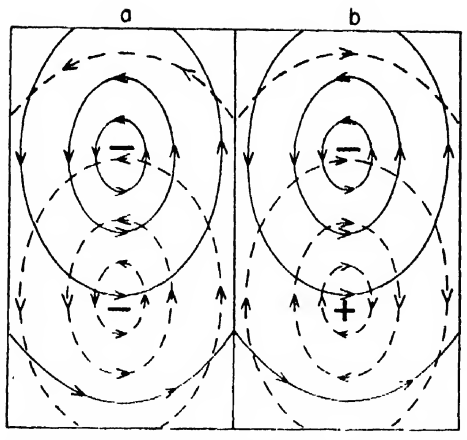


FIG. 7. Streamlines of perturbation centers at 200 mb (a) during in-phase and (b) during out-of-phase superposition of a trough in the westerlies on disturbances in the tropical vortex train aloft.

internal mechanism that can produce the requisite field of divergence.

PROPOSED MODEL OF HURRICANE FORMATION

Analysis of numerous charts at different levels over Atlantic and Pacific Ocean during the hurricane and typhoon season has brought out two factors that indicate the operation of external forces over an area within the tropics where a weak initial disturbance is present in the lower layers.[11]

(1) The trade winds of the low atmosphere disappear quickly with height near 500–400 mb (about 6 km). In the upper troposphere (300–150 mb, 9–15 km) we observe a train of large cyclonic and anticyclonic eddies that gradually drift westward[2] (Fig. 5). These eddies often move at a rate that differs from the displacement of the low level tropical disturbance. The latter, therefore, may be located under a different regime of high level winds at different times. Commonly, disturbances first develop in the air underneath an upper cyclone center. Deepening to typhoon strength, however, often takes place as the surface disturbance is being transferred from the influence of an upper cyclone to that of an upper anticyclone.

We shall choose a position of the deepening disturbance relative to the upper flow shown in Fig. 6 as a model. It is to be noted that the surface center and the air associated with it do not participate in the indicated motion of the current at 200 mb. This northerly current blows over the disturbed zone at rates that attain 50 m.p.h. and more.

(2) There is pronounced interaction at high levels between the tropical vortex train mentioned and the long waves in the westerlies outside the tropics. The low latitude eddies drift westward, and the troughs and ridges in the westerlies drift eastward. At times an upper cyclonic center of the tropics, for example, will be situated south of a trough in the westerlies, and at other times an anticyclonic center. Thus the connection between high level disturbances of middle and low latitudes will vary between in-phase and out-of-phase superposition. In Fig. 7 two types of superposition are sketched. Both disturbances have the same sense of rotation in Fig. 7a, but opposite rotation in Fig. 7b. The north-south components of motion are in the same sense in Fig. 7a, but the zonal component between the perturbation centers is opposite. In Fig. 7b, the reverse holds. Thus the amplitude of the composite streamlines and the intensity of the meridional flow will be increased by the superposition sketched in Fig. 7a, but the zonal speed in the central latitude band will decrease. Conversely, the meridional flow weakens in Fig. 7b and the streamline amplitude becomes flatter, while the zonal component in the central portion speeds up.

Without rigorous analysis we cannot know whether superposition produces net changes of kinetic energy. It is possible that the increase of north-south kinetic energy in case of Fig. 7a is obtained entirely through transformation of east-west into north-south kinetic energy. Our requirements are met if the speed of the northerly current shown in Fig. 6 increases in the region of the incipient tropical disturbance, irrespective of the net kinetic energy change over the whole system of middle and low latitude disturbances aloft. As just seen this occurs during in-phase superposition (cf. also Fig. 5), which therefore is requisite for the proposed model.

The Initial Pressure Fall

The effect of the superposition on the pressure field may be deduced from the equations of horizontal motion on the rotating earth

$$\dot{v}_n + v_s(kv + f - \dot{\alpha}) = -1/\rho(\partial p/\partial n), \qquad (1)$$

$$\dot{v}_s - v_n(kv + f - \dot{\alpha}) = 0. \qquad (2)$$

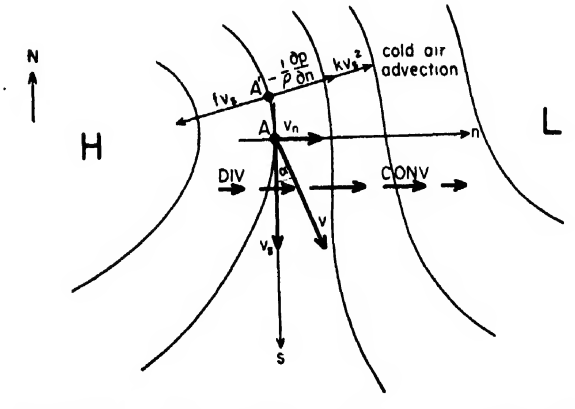


FIG. 8. Illustration of Eqs. (1), (2) at A and of Eq. (4) at A'. H denotes high pressure (resp. high elevation of 200 mb surface) and L low pressure (resp. low elevation of 200 mb surface).

[11] For a discussion of the formation of initial disturbances cf. reference 1.

Here, the coordinates s and n are chosen to lie parallel and normal (toward lower pressure) to the upper isobars (Fig. 8). v is the total motion, v_s the motion along and v_n the motion normal to the isobars (cross-stream circulation); f is the Coriolis parameter, p the pressure, ρ the density, k the trajectory curvature taken positive for cyclonic flow and α the angle between isobars and total motion (Fig. 8). The dots denote substantial differentiation with respect to time. Combination of (1) and (2) yields

$$\frac{1}{2}\frac{dv^2}{dt} = -\frac{v_n}{\rho}\frac{\partial p}{\partial n}, \tag{3}$$

which shows explicitly the well-known relation between acceleration, pressure gradient force and cross-stream flow. If the latter is zero prior to superposition, Eqs. (2) and (3) vanish and (1) becomes the gradient wind equation

$$kv_s^2 + fv_s = -1/\rho(\partial p/\partial n), \tag{4}$$

where $v_s = v$ is the total motion. In case of anticyclonic flow the centrifugal force kv_s^2 and the pressure gradient force $-1/\rho(\partial p/\partial n)$ act to the left of the isobars looking downstream; the Coriolis force fv_s acts to the right (Fig. 8).

We now let the superposition take place, manifested by an increase of $-1/\rho(\partial p/\partial n)$. The Coriolis force fv_s no longer affords balance. A cross-isobar flow v_n toward lower pressure is indicated and also an increase of the total motion (kinetic energy). Figure 6 shows the streamlines of the total motion while the cross-stream flow is in progress.

On account of the motion normal to the isobars mass divergence takes place to the right of the northerly current looking downstream, and mass convergence to its left (Fig. 8). In the region of divergence the surface pressure falls because of the mass removal, and the initial surface disturbance begins to deepen. This is the desired result, and it is brought about entirely by a coincidence: presence of the surface depression under the northerly current at the time that in-phase superposition develops. Otherwise the superposition—an event observed in several areas of the globe on every day—has no particular consequences for reasons to appear later. Nor does the existence of a weak low-level depression in a position as sketched in Fig. 6 warrant a forecast of typhoon formation per se.

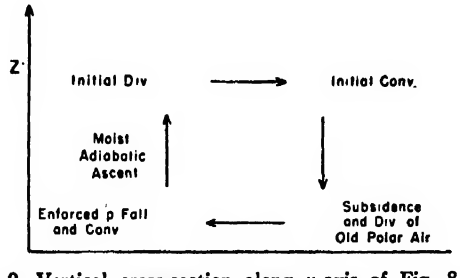


FIG. 9. Vertical cross-section along n-axis of Fig. 8. Model illustrates hypothetical development of vertical cross-stream circulation leading to deepening of the surface depression.

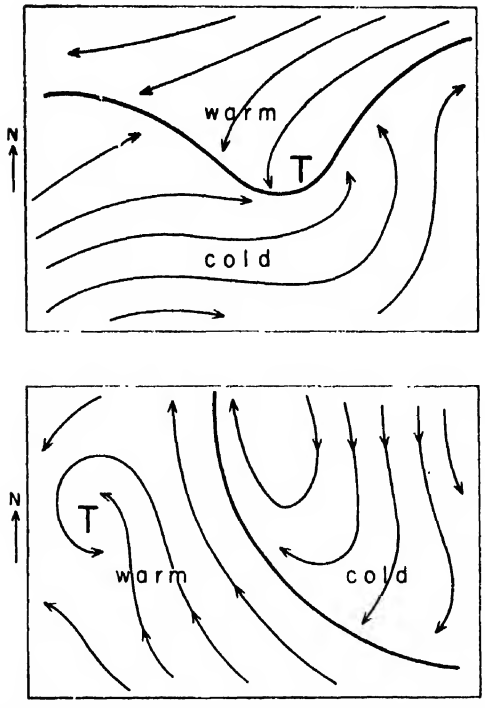


FIG. 10. Classical model of typhoon formation at "equatorial front" (above) and proposed revision (below) with cyclonic vortex situated entirely in the warm air.

Cross-Stream Circulation

The generation of the high level cross-stream flow and attendant field of convergence and divergence has several obvious consequences. Surface convergence develops under the region of upper divergence and, conversely, surface divergence develops under the region of upper convergence. It follows from continuity requirements that air must ascend along the right margin of the northerly current aloft and descend along its left margin (looking downstream). Figure 9 shows the complete cross-stream circulation in vertical cross-section along the n-axis, which points from high toward low pressure aloft. This circulation acts as a left-handed screw about the s-axis. The observations show that temperatures are higher on the right of this axis than on its left (Fig. 2). Therefore the warm air rises and the cold air sinks. The vertical circulation is in the kinetic energy producing sense.

We can now make use of another fact supplied by the observations. In the situation described a strong equatorward acceleration of the low level northeast trade frequently is noted along the eastern end of the upper northerly current and still farther east (500–1000 km from the tropical depression). Sometimes this acceleration is coupled with arrival of a polar air mass from the north.[3] Such polar air masses are gradually transformed

to tropical air during their equatorward passage. They are also observed to spread laterally and therefore sink. This sinking of old polar air adds to the descent in the eastern portion of the circulation cell shown in Fig. 9 and therefore acts to accelerate the whole cell. The polar air, of course, cannot invade the area of the tropical depression itself, if there is to be a storm. It bypasses this area going southward (Fig. 10). If, for some reason, the polar outbreak takes place too far to the west, the whole development ceases.

It is of interest that the model here proposed does find a role for old polar air, so prominently mentioned in many previous studies. This role differs radically from that formerly assigned to the polar air. Figure 10 illustrates the difference in viewpoint on the relative position of polar air and tropical disturbance. According to the present model, the sinking of the accelerated northeast trade provides a second mechanism that is suitable to remove air aloft from the area of the tropical depression and allow it to settle at a considerable distance. It is to be noted that this particular mechanism is not encountered in all instances. Among several alternatives, the formation of clockwise eddies in the low latitude trade stream has been mentioned.[2]

Maintenance of Cross-Stream Circulation

It is necessary now to examine the crucial problem of maintenance of the vertical circulation cell that has been initiated. There are two principal factors that act to restore equilibrium as soon as it is disturbed. One of these is the Coriolis force; the other the vertical stratification of the atmosphere. Except under special assumptions—which are often unrealistic—Eqs. (1, 2) do not yield a unique answer as there are too many variables. Consider, for instance, a situation near latitude 20° when initially $v_s = 10$ m sec.$^{-1}$ and $k = 10^{-6}$ m^{-1} (radius of curvature of 1000 km). Then $fv_s = 4 \times 10^{-2}$ cm sec.$^{-2}$ and $kv_s^2 = 10^{-2}$ cm sec.$^{-2}$. Let the windspeed v double in consequence of changes in the pressure gradient force Eq. (3). As $v_n \ll v_s$ in most situations of the kind here discussed, we can still estimate the magnitude of the forces in Eq. (1) by setting $v_s \sim v$. Both kv_s^2 and fv_s increase by 4×10^{-2} cm sec.$^{-2}$ provided k remains constant. In this case a true dynamic instability would have been achieved, especially as the term $-\dot{\alpha}v_s$, whose magnitude is hard to judge, acts in the same direction as the centrifugal force. The increase of restoring force is balanced by the increase of the centrifugal force. Thus the initial increment of the pressure gradient force remains unbalanced. As k is small, however, large percentual changes of this quantity are possible. Moreover, the cross-stream circulation affects the mass distribution and therewith the pressure gradient force. Only one definite statement is possible. A cross-stream circulation, once initiated, is much more easily maintained when the flow is anticyclonically curved than when it is cyclonically curved. In the latter case, fv_s and kv_s^2 act together as restoring forces.

The energy aspects yield a more conclusive answer. Let us return to curve (a) of Fig. 2 which serves as a representative sample of non-saturated air as encountered in the tropics. When such air is raised following the dry-adiabatic path indicated by the horizontal lines of Fig. 2, cooling results at all heights. When it sinks, warming takes place. Assuming that the vertical displacements sketched in Fig. 9 occur in non-saturated air, it follows that the western portion of the circulation cell will cool and that the eastern portion will become warmer. This acts to destroy the initial temperature gradient between eastern and western cell ends. It also acts to lower the pressure at high levels in the west and raise it in the east, reducing the increment in the pressure gradient force aloft that had been obtained through superposition. Clearly, the vertical circulation cell contains an internal mechanism for its own destruction under the circumstances described. The order of magnitude of vertical displacement requisite to stop the circulation is small. It may be estimated at 100 m.

It is quite probable that most vertical circulations actually come to an end in this way. If there is to be a hurricane, further necessary conditions must be introduced. At the eastern cell end, the temperature field relative to the incipient disturbance can be maintained if the warming due to subsidence is balanced by continuous advection of colder air. Circumstances are favorable for this to occur. The frequent equatorward movement of an old polar airmass in the low levels 500–1000 km east of a developing disturbance has already been mentioned. *At high levels, the superposition and attendant increase of streamline amplitude make possible prolonged displacement of cooler air from middle latitudes toward the tropics.* The data support this suggestion. Observations in a number of representative cases have shown that temperatures remained constant, or even decreased, in the eastern parts of the vertical circulation cells, in spite of subsidence.

In the region of ascent, temperature falls can be prevented by condensation. If the air that forms the tropical depression in the lower layers is saturated (area 10^3–10^4 sq. km), its ascent follows the moist rather than the dry-adiabatic path and no cooling will occur. *Thus it is essential to have such a saturated air mass available.* The saturation is produced by formation and dissipation of numerous cumulus clouds in the disturbed area. Moisture is funnelled upward by this mechanism and then spread laterally aloft. It is of importance to note that this process takes place mainly below six km. As stated initially, a rapid transition from trade wind to upper vortex regime is observed near that level. The lower air columns have only a slow drift relative to the surface depression. Whereas the air at 200 mb can pass over it in perhaps three hours, the low level air may remain in its orbit for days. Thus the moisture that is funnelled aloft,

is not carried away and gradual saturation below the level of wind reversal is made possible.

It is of much interest to note that such penetration of moisture is known to coincide with a gradual approach of the temperature lapse rate to the moist adiabatic. When saturation is complete, the vertical stratification of the atmosphere is nearly neutral and no longer unstable with respect to moist adiabatic displacements of particles. The local buoyancy force becomes relatively unimportant. This observation reveals another difference between dynamic and thermal theories of hurricane formation. The latter wish to obtain energy for the development of a circulation from vertical instability with respect to moist-adiabatic ascent. For practical purposes, this means that the atmosphere is unsaturated except for the rising parcels. As just seen, this is precisely the situation in which a large-scale vertical circulation cell will not be maintained. Of course, it would be best to operate with a field of upper divergence on a lower atmosphere that is saturated and has a temperature lapse rate in excess of the moist adiabatic. Since this cannot be obtained in reality, *an atmosphere as given by curve (b) of Fig. 2 represents optimum conditions for cyclone formation. Condensation energy is converted into kinetic energy not through local overturning of a vertically unstable atmosphere, but through large-scale vertical circulations. It follows that any mechanism, applied to the lower atmosphere, that operates to establish curve (b) of Fig. 2 will augment the chances of hurricane development.*

Conclusion

The foregoing pages have attempted to determine the reasons why tropical storms form under certain circumstances. In particular, we have sought to establish a mechanism for the removal of air over a cyclogenetic zone at high levels. This led to an inversion of previous attacks on the problem. Former investigators at first tried to find reasons for an increase of the cyclonic circulation from low-level considerations and then looked for ways and means to remove the air ascending in the vortex core. Such an approach restricts the analysis to internal factors—latent heat of condensation and internal structure of the field of motion.

In this report, the mechanism for outflow is established prior to deepening, and the cyclone is considered to deepen in response to upper divergence. A total of three energy sources has been applied to this end. The in-phase superposition of high and low latitude wave trains in the upper troposphere directly intensifies the speed of the high level northerly current that has been analyzed under the assumption that the balance of forces within it is the decisive factor. Subsidence of a southward moving mass of polar air provides a second energy source underneath the left-hand portion of this current, and the latent heat of condensation acts as third source on its right.

As yet, atmospheric processes cannot be measured directly but must be inferred from the three-dimensional structure of the atmosphere and its changes with time. These processes are not tied uniquely to the configurations of upper or lower flow pattern. We know, for instance, that it is not common to have ascent with high level divergence in the eastern portions of dynamic highs. Most of the time convergence and descent prevail aloft in these areas. Again, hurricanes have been observed to deepen when passing underneath upper troughs. But the shape of these troughs and the temperature distribution within them are not unique. Streamline and/or trajectory curvature can be cyclonic or anticyclonic.

Certain basic features of the *processes* that lead to cyclone formation must be alike if the same result is to be achieved at the surface. This study has attempted to uncover some of these processes. It is a task for the future to analyze other types of flow pattern configuration that lead to cyclogenesis.

ACKNOWLEDGMENT

The writer is greatly indebted to Dr. H. R. Byers, who took much time in helping to organize the material of this report. Drs. J. Bjerknes and E. Palmén discussed several aspects of the problems with the writer, and some phases of the study were undertaken following these discussions. Acknowledgment also is due to Dr. T. C. Yeh and Mr. N. La Seur, who participated in the analysis of many aspects of the topics treated.

Nuclei of Strain in the Semi-Infinite Solid

RAYMOND D. MINDLIN AND DAVID H. CHENG
Columbia University, New York, New York
(Received April 28, 1950)

Stress functions are given for forty nuclei of strain in the semi-infinite elastic solid. The corresponding stresses have the property that the traction vanishes across the plane boundary of the solid. The nuclei are derived, by processes of superposition, differentiation and integration, from the solution of the elasticity equations for the single force in the interior of the semi-infinite solid.

I. INTRODUCTION

THERE is an interesting group of solutions of the equations of the linear theory of elasticity known as nuclei of strain. For the isotropic solid of indefinite extent the fundamental solution of this type is that of Kelvin[1] for a single force applied at a point. By differentiation of this solution a family of additional nuclei can be obtained, examples of which are[1] the double force and the double force with moment. These, in turn, can be combined to form nuclei such as the center of dilatation and the center of rotation. By further differentiation, various types of doublets and multiplets are obtained. Alternatively, the preceding solutions may be integrated to form line distributions, such as[1] semi-infinite lines of centers of dilatation, or doublets, or double forces with moment, etc.

The solutions of many interesting problems may be expressed in terms of combinations of these nuclei or have been discovered by a process of superposition of nuclei. Examples are Lamé's problem[2] of the spherical container under internal and external pressure, Boussinesq's and Cerruti's problems[3] of the single force acting at the surface of a semi-infinite solid, Southwell's problem[2] of the spherical cavity in an infinite solid under simple tension, the analogous solution by Goodier[4] for the spherical inclusion and Mindlin's solution[5] for the single force in the semi-infinite solid.

The last named solution occupies the same position for the semi-infinite solid as Kelvin's solution does for the solid of indefinite extent. The nuclei derived from it, by differentiation and integration, all satisfy the condition of vanishing traction on a plane boundary so that, in contemplating the solution of new problems with at least one plane, traction-free surface, a portion of the boundary conditions is already satisfied.

Since the solution for the single force in the semi-infinite solid is itself formed from the superposition of eighteen nuclei derived from Kelvin's solution (six for each of the three components of the force), the calculation of nuclei for the semi-infinite solid is a formidable one. The possible usefulness of such nuclei is so great, however, that this task was undertaken and the results for forty nuclei are reported below.[6] The methods employed are the same as those described by Love[1] in connection with the calculation of nuclei in the solid of indefinite extent. The results are given in very condensed form by expressing them in terms of the Galerkin vector stress function. The displacement vector $\mathbf{u}$ is given in terms of the Galerkin vector $\mathbf{F}$ by

$$2G\mathbf{u}=2(1-\nu)\nabla^2\mathbf{F}-\nabla\nabla\cdot\mathbf{F}, \tag{1}$$

where G is the shear modulus, ν is Poisson's ratio and ∇, $\nabla\cdot$ and ∇^2 are the gradient, divergence and Laplacian operators, respectively.

II. REDUCTION FROM THREE COMPONENTS TO ONE

Certain nuclei are obtained by the superposition of three others and the immediate result is a three-component Galerkin vector. This may be simplified, in some cases, by reduction to a one-component vector, say $\mathbf{F}'=\mathbf{k}Z'$, so that

$$2G\mathbf{u}=2(1-\nu)\mathbf{k}\nabla^2Z'-\nabla(\partial Z'/\partial z). \tag{2}$$

To find the function $\mathbf{F}'$ corresponding to a given $\mathbf{F}$, (1) and (2) are equated to obtain the differential equation

$$2(1-\nu)\mathbf{k}\nabla^2Z'-\nabla(\partial Z'/\partial z)=2(1-\nu)\nabla^2\mathbf{F}-\nabla\nabla\cdot\mathbf{F}, \tag{3}$$

which is to be solved for Z' in terms of $\mathbf{F}$.

Case 1: $\nabla^2\mathbf{F}=0$

Operating on (3) with $\nabla\cdot$ and $\nabla\times$ and using the condition $\nabla^2\mathbf{F}=0$, there results

$$\nabla\nabla^2Z'=0$$

from which ∇^2Z' is a constant which may be taken as zero since, by (2), it contributes only a rigid body displacement. Hence (3) reduces to

$$\nabla(\partial Z'/\partial z)=\nabla\nabla\cdot\mathbf{F} \tag{4}$$

the first integral of which is

$$\partial Z'/\partial z=\nabla\cdot\mathbf{F}$$

[1] A. E. H. Love, *Mathematical Theory of Elasticity* (Cambridge University Press, London, 1927), fourth edition, Chapter VIII.
[2] S. Timoshenko, *Theory of Elasticity* (McGraw-Hill Book Company, Inc., New York, 1934), Chapter 11.
[3] See reference 1, Chapter X.
[4] J. N. Goodier, "Concentration of stress around spherical and cylindrical inclusions and flaws," J. App. Mech. 1, 39–44 (1933).
[5] R. D. Mindlin, "Force at a point in the interior of a semi-infinite solid," Physics 7, 195–202 (1936).
[6] An application of one of them is given in J. App. Phys. 21, 931 (1950).

plus a constant which may be omitted since it does not contribute to the displacement. The second integral of (4) is

$$Z'=\int \nabla\cdot\mathbf{F}dz$$

plus a function of x and y which may be omitted since it does not contribute to the displacement.

Case 2: $\nabla\times\nabla\times\mathbf{F}=\nabla^2\nabla\cdot\mathbf{F}=0$

Since

$$\nabla\times\nabla\times\mathbf{F}=\nabla\nabla\cdot\mathbf{F}-\nabla^2\mathbf{F},$$

Eq. (3) reduces, in this case, to

$$2(1-\nu)\mathbf{k}\nabla^2Z'-\nabla(\partial Z'/\partial z)=(1-2\nu)\nabla\nabla\cdot\mathbf{F}. \quad (5)$$

Operating on (5) with $\nabla\cdot$ and $\nabla\times$ and using the condition $\nabla^2\nabla\cdot\mathbf{F}=0$, we find, as before, that $\nabla^2Z'=0$. Hence (5) reduces to

$$\nabla(\partial Z'/\partial z)=-(1-2\nu)\nabla\nabla\cdot\mathbf{F}$$

from which, by the procedure used in Case 1,

$$Z'=-(1-2\nu)\int\nabla\cdot\mathbf{F}dz.$$

III. HALF-SPACE NUCLEI IN TERMS OF GALERKIN VECTORS

The half-space considered is defined by $Z\geqslant 0$. Point nuclei are at $(0, 0, c)$ and one end of each line nucleus is at the same point. $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are unit vectors in the x, y, z directions, respectively, and

$$R_1^2=x^2+y^2+(z-c)^2, \quad R_2^2=x^2+y^2+(z+c)^2.$$

A multiplying constant is omitted in each stress function.

A. Single Force

(1) Single force in x direction

$$\mathbf{i}\left\{R_1+R_2-\frac{2c^2}{R_2}+4(1-\nu)(1-2\nu)\times[(z+c)\log(R_2+z+c)-R_2]\right\}+\mathbf{k}\left[\frac{2cx}{R_2}+2(1-2\nu)x\log(R_2+z+c)\right].$$

(2) Single force in y direction

$$\mathbf{j}\left\{R_1+R_2-\frac{2c^2}{R_2}+4(1-\nu)(1-2\nu)\times[(z+c)\log(R_2+z+c)-R_2]\right\}+\mathbf{k}\left[\frac{2cy}{R_2}+2(1-2\nu)y\log(R_2+z+c)\right]$$

(3) Single force in z direction

$$\mathbf{k}\left\{R_1+[8\nu(1-\nu)-1]R_2+4(1-2\nu)[(1-\nu)z-\nu c]\log(R_2+z+c)-\frac{2cz}{R_2}\right\}.$$

B. Double Force

(1) Double force in x direction

$$\mathbf{i}x\left[\frac{1}{R_1}+\frac{1}{R_2}-\frac{4(1-\nu)(1-2\nu)}{R_2+z+c}+\frac{2c^2}{R_2^3}\right]+\mathbf{k}\left[\frac{2c}{R_2}-\frac{2cx^2}{R_2^3}+\frac{2(1-2\nu)x^2}{R_2(R_2+z+c)}+2(1-2\nu)\log(R_2+z+c)\right].$$

(2) Double force in y direction

$$\mathbf{j}y\left[\frac{1}{R_1}+\frac{1}{R_2}-\frac{4(1-\nu)(1-2\nu)}{R_2+z+c}+\frac{2c^2}{R_2^3}\right]+\mathbf{k}\left[\frac{2c}{R_2}+2(1-2\nu)\log(R_2+z+c)+\frac{2(1-2\nu)y^2}{R_2(R_2+z+c)}-\frac{2cy^2}{R_2^3}\right].$$

(3) Double force in z direction

$$\mathbf{k}\left[\frac{z-c}{R_1}-(1-4\nu)\frac{z-c}{R_2}-\frac{2cz(z+c)}{R_2^3}+4\nu(1-2\nu)\log(R_2+z+c)\right].$$

C. Double Force with Moment

(1) Double force in x direction with moment about y axis

$$\mathbf{i}\left[\frac{z-c}{R_1}-\frac{z-3c}{R_2}-\frac{2c^2(z+c)}{R_2^3}-4(1-\nu)(1-2\nu)\log(R_2+z+c)\right]+\mathbf{k}2x\left[\frac{c(z+c)}{R_2^3}-\frac{2(1-\nu)}{R_2}\right].$$

(2) Double force in x direction with moment about z axis

$$\mathbf{i}y\left[\frac{1}{R_1}+\frac{1}{R_2}+\frac{2c^2}{R_2^3}-\frac{4(1-\nu)(1-2\nu)}{R_2+z+c}\right]+\mathbf{k}2xy\left[\frac{1-2\nu}{R_2(R_2+z+c)}-\frac{c}{R_2^3}\right].$$

(3) Double force in y direction with moment about x axis

$$\mathbf{j}\left[\frac{z-c}{R_1}-\frac{z-3c}{R_2}-\frac{2c^2(z+c)}{R_2^3}-4(1-\nu)(1-2\nu)\log(R_2+z+c)\right]+\mathbf{k}2y\left[\frac{c(z+c)}{R_2^3}-\frac{2(1-\nu)}{R_2}\right].$$

(4) Double force in y direction with moment about z axis

$$\mathbf{j}x\left[\frac{1}{R_1}+\frac{1}{R_2}+\frac{2c^2}{R_2^3}-\frac{4(1-\nu)(1-2\nu)}{R_2+z+c}\right]+\mathbf{k}2xy\left[\frac{1-2\nu}{R_2(R_2+z+c)}-\frac{c}{R_2^3}\right].$$

(5) Double force in z direction with moment about x axis

$$\mathbf{k}y\left\{\frac{1}{R_1}+\frac{8\nu(1-\nu)-1}{R_2}+\frac{4(1-2\nu)[(1-\nu)z-\nu c]}{R_2(R_2+z+c)}+\frac{2cz}{R_2^3}\right\}.$$

(6) Double force in z direction with moment about y axis

$$\mathbf{k}x\left\{\frac{1}{R_1}+\frac{8\nu(1-\nu)-1}{R_2}+\frac{4(1-2\nu)[(1-\nu)z-\nu c]}{R_2(R_2+z+c)}+\frac{2cz}{R_2^3}\right\}.$$

D. Line of Double Forces with Moment

(1) Type C(1) along x axis from $x=0$ to $x=-\infty$

$$\mathbf{i}\left[(z-c)\log(R_1+x)-(z-3c)\log(R_2+x)-\frac{2c^2(z+c)x}{R_2(R_2^2-x^2)}\right]-\mathbf{k}\left[\frac{2c(z+c)}{R_2}+8\nu(1-\nu)R_2+4(1-\nu)(1-2\nu)(z+c)\log(R_2+z+c)\right].$$

(2) Type C(2) along x axis from $x=0$ to $x=-\infty$

$$\mathbf{i}y\left[\log(R_1+x)+\log(R_2+x)+\frac{2c^2x}{R_2(R_2^2-x^2)}\right]+\mathbf{k}y\left[2\nu(1-2\nu)\log(R_2+z+c)-2(1-\nu)(1-2\nu)\frac{z+c}{R_2+z+c}+\frac{2c}{R_2}\right].$$

(3) Type C(3) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{j}\left[(z-c)\log(R_1+y)-(z-3c)\log(R_2+y)-\frac{2c^2(z+c)y}{R_2(R_2^2-y^2)}\right]-\mathbf{k}\left[\frac{2c(z+c)}{R_2}+8\nu(1-\nu)R_2+4(1-\nu)(1-2\nu)(z+c)\log(R_2+z+c)\right].$$

(4) Type C(4) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{j}x\left[\log(R_1+y)+\log(R_2+y)+\frac{2c^2y}{R_2(R_2^2-y^2)}\right]+\mathbf{k}x\left[2\nu(1-2\nu)\log(R_2+z+c)-2(1-\nu)(1-2\nu)\frac{z+c}{R_2+z+c}+\frac{2c}{R_2}\right].$$

(5) Type C(5) along z axis from $z=c$ to $z=\infty$

$$\mathbf{k}y\left\{\log(R_1-z+c)+[8\nu(1-\nu)-1]\log(R_2+z+c)-\frac{4(1-2\nu)z}{R_2+z+c}-2\nu(1-2\nu)\left[\log(R_2+z+c)+\frac{R_2}{R_2+z+c}\right]-2z\left[\frac{1}{R_2}+\frac{z(z+c)}{R_2(x^2+y^2)}-\frac{z}{(x^2+y^2)}\right]\right\}.$$

(6) Type C(6) along z axis from $z=c$ to $z=\infty$

$$\mathbf{k}x\left\{\log(R_1-z+c)+[8\nu(1-\nu)-1]\log(R_2+z+c)-\frac{4(1-2\nu)z}{R_2+z+c}-2\nu(1-2\nu)\left[\log(R_2+z+c)+\frac{R_2}{R_2+z+c}\right]-2z\left[\frac{1}{R_2}+\frac{z(z+c)}{R_2(x^2+y^2)}-\frac{z}{(x^2+y^2)}\right]\right\}.$$

E. Center of Dilatation

(1)
$$\mathbf{i}[\log(R_1+x)+(1-4\nu)\log(R_2+x)]+\mathbf{k}\frac{2z}{R_2}.$$

(2)
$$\mathbf{j}[\log(R_1+y)+(1-4\nu)\log(R_2+y)]+\mathbf{k}\frac{2z}{R_2}.$$

(3)
$$\mathbf{k}\left[\log(R_1+z-c)+(1-4\nu)\log(R_2+z+c)+\frac{2z}{R_2}\right].$$

All three yield the same displacements and stresses.

F. Line of Centers of Dilatation

(1) Along x axis from $x=0$ to $x=-\infty$

$$\mathbf{i}\{x\log(R_1+x)-R_1+(1-4\nu)[x\log(R_2+x)-R_2]\}+\mathbf{k}2z\log(R_2+x).$$

(2) Along y axis from $y=0$ to $y=-\infty$

$$\mathbf{j}\{y\log(R_1+y)-R_1+(1-4\nu)[y\log(R_2+y)-R_2]\}+\mathbf{k}2z\log(R_2+y).$$

(3) Along z axis from $z=c$ to $z=\infty$

$$\mathbf{k}\{R_1-(z-c)\log(R_1+z-c)+(1-4\nu)\times[(z+c)\log(R_2+z+c)-R_2]+2z\log(R_2+z+c)\}.$$

G. Doublet

(1) Axis of doublet parallel to x axis

$$\mathbf{i}\left(\frac{1}{R_1}+\frac{1-4\nu}{R_2}\right)-\mathbf{k}\frac{2xz}{R_2^3}.$$

(2) Axis of doublet parallel to y axis

$$\mathbf{j}\left(\frac{1}{R_1}+\frac{1-4\nu}{R_2}\right)-\mathbf{k}\frac{2yz}{R_2^3}.$$

(3) Axis of doublet parallel to z axis

$$\mathbf{k}\left[\frac{1}{R_1}-\frac{1-4\nu}{R_2}+\frac{2z(z+c)}{R_2^3}\right].$$

H. Linearly Varying Line of Doublets with Strength Proportional to Distance from the Origin

(1) Type G(1) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{i}\{R_1-y\log(R_1+y)+(1-4\nu)[R_2-y\log(R_2+y)]\}+\mathbf{k}\frac{2xz}{R_2+y}.$$

(2) Type G(1) along z axis from $z=c$ to $z=\infty$

$$\mathbf{i}\{R_1+z\log(R_1-z+c)+(1-4\nu)[R_2-z\log(R_2+z+c)]\}+\mathbf{k}2xz\left[\frac{1}{R_2}+\frac{z(z+c)}{R_2(x^2+y^2)}-\frac{z}{x^2+y^2}\right].$$

(3) Type G(2) along x axis from $x=0$ to $x=-\infty$

$$\mathbf{j}\{R_1-x\log(R_1+x)+(1-4\nu)[R_2-x\log(R_2+x)]\}+\mathbf{k}\frac{2yz}{R_2+x}.$$

(4) Type G(2) along z axis from $z=c$ to $z=\infty$

$$\mathbf{j}\{R_1+z\log(R_1-z+c)+(1-4\nu)[R_2-z\log(R_2+z+c)]\}+\mathbf{k}2yz\left[\frac{1}{R_2}+\frac{z(z+c)}{R_2(x^2+y^2)}-\frac{z}{x^2+y^2}\right].$$

(5) Type G(3) along x axis from $x=0$ to $x=-\infty$

$$\mathbf{k}\left\{R_1-(1-4\nu)R_2-x[\log(R_1+x)-(1-4\nu)\log(R_2+x)]-\frac{2z(z+c)}{R_2+x}\right\}.$$

(6) Type G(3) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{k}\left\{R_1-(1-4\nu)R_2-y[\log(R_1+y)-(1-4\nu)\log(R_2+y)]-\frac{2z(z+c)}{R_2+y}\right\}.$$

I. Center of Rotation

(1) Center of rotation about x axis (Type C(5)—Type C(3))

$$\mathbf{j}\left[-\frac{z-c}{R_1}+\frac{z-3c}{R_2}+4(1-\nu)(1-2\nu)\log(R_2+z+c)+\frac{2c^2(z+c)}{R_2^3}\right]+\mathbf{k}y\left\{\frac{1}{R_1}+\frac{4\nu(1-2\nu)+3}{R_2}-\frac{2c^2}{R_2^3}+\frac{4(1-2\nu)[(1-\nu)z-\nu c]}{R_2(R_2+z+c)}\right\}.$$

(2) Center of rotation about y axis (Type C(6)—Type C(1))

$$\mathbf{i}\left[-\frac{z-c}{R_1}+\frac{z-3c}{R_2}+4(1-\nu)(1-2\nu)\log(R_2+z+c)+\frac{2c^2(z+c)}{R_2^3}\right]+\mathbf{k}x\left\{\frac{1}{R_1}+\frac{4\nu(1-2\nu)+3}{R_2}-\frac{2c^2}{R_2^3}+\frac{4(1-2\nu)[(1-\nu)z-\nu c]}{R_2(R_2+z+c)}\right\}.$$

(3) Center of rotation about z axis (Type C(4)—Type C(2))

$$(\mathbf{j}x-\mathbf{i}y)\left[\frac{1}{R_1}+\frac{1}{R_2}+\frac{2c^2}{R_2^3}-\frac{4(1-\nu)(1-2\nu)}{R_2+z+c}\right].$$

J. Line of Centers of Rotation

(1) Type I(1) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{j}\left[-(z-c)\log(R_1+y)+(z-3c)\log(R_2+y)+\frac{2c^2(z+c)y}{R_2(R_2^2-y^2)}\right]+\mathbf{k}\left\{R_1-[4(1-2\nu)^2-3]R_2+4(1-2\nu)[z+(1-2\nu)(z+c)]\log(R_2+z+c)+\frac{2c^2}{R_2}\right\}.$$

(2) Type I(1) along z axis from $z=c$ to $z=\infty$

$$\mathbf{j}\left\{R_1+R_2-\frac{2c^2}{R_2}+4(1-\nu)(1-2\nu)\times[(z+c)\log(R_2+z+c)-R_2]\right\}+\mathbf{k}y\left\{\log(R_1-z+c)+[2\nu(1-2\nu)+1]\log(R_2+z+c)+\frac{2c^2}{R_2(R_2+z+c)}-\frac{2[R_2(z-c)-z^2]}{x^2+y^2}+\frac{2(1-2\nu)}{R_2+z+c}[\nu R_2+2\nu(z+c)-2z]\right\}.$$

(3) Type I(2) along x axis from $x=0$ to $x=-\infty$

$$\mathbf{i}\left[-(z-c)\log(R_1+x)+(z-3c)\log(R_2+x)+\frac{2c^2(z+c)x}{R_2(R_2^2-x^2)}\right]+\mathbf{k}\left\{R_1-[4(1-2\nu)^2-3]R_2+4(1-2\nu)[z+(1-2\nu)(z+c)]\log(R_2+z+c)+\frac{2c^2}{R_2}\right\}.$$

(4) Type I(2) along z axis from $z=c$ to $z=\infty$

$$\mathbf{i}\left\{R_1+R_2-\frac{2c^2}{R_2}+4(1-\nu)(1-2\nu)\times[(z+c)\log(R_2+z+c)-R_2]\right\}+\mathbf{k}x\left\{\log(R_1-z+c)+[2\nu(1-2\nu)+1]\log(R_2+z+c)+\frac{2c^2}{R_2(R_2+z+c)}-\frac{2[R_2(z-c)-z^2]}{x^2+y^2}+\frac{2(1-2\nu)}{R_2+z+c}[\nu R_2+2\nu(z+c)-2z]\right\}.$$

(5) Type I(3) along x axis from $x=0$ to $x=-\infty$

$$-\mathbf{i}y\left[\log(R_1+x)+\log(R_2+x)+\frac{2c^2x}{R_2(R_2^2-x^2)}\right]+\mathbf{j}\left[R_1+R_2-\frac{2c^2}{R_2}\right].$$

(6) Type I(3) along y axis from $y=0$ to $y=-\infty$

$$\mathbf{i}\left[R_1+R_2-\frac{2c^2}{R_2}\right]-\mathbf{j}x\left[\log(R_1+y)+\log(R_2+y)+\frac{2c^2y}{R_2(R_2^2-y^2)}\right].$$

IV. COMPONENTS OF DISPLACEMENT AND STRESS

The formulas for the components of displacement and stress corresponding to each of the nuclei listed in the preceding section are listed in Appendices I and II.[7]

[7] On file at the Department of Civil Engineering, Columbia University, New York 27, New York.

Thermoelastic Stress in the Semi-Infinite Solid

RAYMOND D. MINDLIN AND DAVID H. CHENG
Columbia University, New York, New York
(Received April 28, 1950)

The solution of the elasticity equations for the center of dilatation in the semi-infinite solid is introduced into Goodier's theory of thermoelastic stress. It is shown that the problem of potential to be solved is identical with that for the solid of indefinite extent. The results are applied to the case of an expanding (or contracting) spherical inclusion embedded in an elastic body near its surface.

I. INTRODUCTION

GOODIER[1] has shown that the elastic effect of a non-uniform distribution of temperature (T) in an infinite body is the same as that of a distribution of centers of dilatation of strength $-\beta/4\pi$ where

$$\beta=\alpha T(1+\nu)/(1-\nu) \tag{1}$$

and α and ν are the coefficient of linear thermal expansion and Poisson's ratio, respectively. Since α and T always appear together, what is said of a non-uniform distribution of temperature applies also to a space variation of the coefficient of linear thermal expansion. Goodier's theory has been applied to two problems of thermal stress in the infinite elastic body by Myklestad.[2]

If the body is not infinite the singularities within it give rise to tractions and displacements at its surface and the removal of these is a boundary-value problem in elasticity theory. However, Goodier noted that, if the solution of the elasticity equations for the center of dilatation is known for the bounded body, the thermoelastic problem is reduced, again, to an integration if the method of singularities is employed.

In this paper it is shown that, for the semi-infinite body with traction-free surface, the additional integral has the same form as that for the infinite body so that, if the latter is known, the calculation of displacement and stress reduces to a process of differentiation only. The result is then applied to a case of a spherical inclusion in the semi-infinite body. Use is made of the solution, given in an accompanying paper,[3] for the center of dilatation in the semi-infinite body.

II. THERMOELASTIC DISPLACEMENT

The Galerkin vector stress function for a center of dilatation of strength A at (ξ, η, ζ) in the infinite, isotropic, elastic solid is[4]

$$\mathbf{F}=-2AG\mathbf{k}\log(R_1'+z-\zeta),$$

where G is the modulus of rigidity, $\mathbf{k}$ is a unit vector in the z direction and

$$R_1'^2=(x-\xi)^2+(y-\eta)^2+(z-\zeta)^2.$$

Hence Goodier's nucleus of thermoelastic strain is given by

$$\mathbf{F}=(G\beta/2\pi)\mathbf{k}\log(R_1'+z-\zeta).$$

In an accompanying paper[3] it is shown that the corresponding nucleus at (ξ, η, ζ) in the semi-infinite body $z\geqslant 0$ (with the plane $z=0$ free of traction) is represented by

$$\mathbf{F}_0=(G\beta/2\pi)\mathbf{k}[\log(R_1'+z-\zeta)+(1-4\nu)\log(R_2'+z+\zeta)+2(z+\zeta)/R_2'-2\zeta/R_2'],$$

where

$$R_2'^2=(x-\xi)^2+(y-\eta)^2+(z+\zeta)^2.$$

Thus, for each center of dilatation at (ξ, η, ζ) in $z>0$, the plane $z=0$ is freed of traction by applying, at the image point $(\xi, \eta, -\zeta)$, a center of dilatation, a double force and a doublet, each of the proper strength.

The displacement $\mathbf{u}_0$ associated with $\mathbf{F}_0$ is calculated from

$$2G\mathbf{u}_0=2(1-\nu)\nabla^2\mathbf{F}_0-\nabla\nabla\cdot\mathbf{F}_0$$

and may be written in terms of operations on $1/R_1'$ and $1/R_2'$ as follows:

$$4\pi\mathbf{u}_0=-\beta\nabla(1/R_1')-\beta\nabla_2(1/R_2'),$$

where

$$\nabla_2=(3-4\nu)\nabla+2\nabla z(\partial/\partial z)-4(1-\nu)\mathbf{k}\nabla^2 z$$

is a vector operator in which, for example, the scalar operator $\nabla^2 z$ signifies that the operand is first to be multiplied by z and then operated upon by ∇^2.

For a distribution $\beta(\xi, \eta, \zeta)$ in a region V_1 in $z>0$, the displacement is given by

$$\begin{aligned}4\pi\mathbf{u}&=4\pi\int_{V_1}\mathbf{u}_0 d\tau\\&=-\nabla\int_{V_1}\frac{\beta}{R_1'}d\tau-\nabla_2\int_{V_1}\frac{\beta}{R_2'}d\tau\\&=-\nabla\int_{V_1}\frac{\beta}{R_1'}d\tau-\nabla_2\int_{V_2}\frac{\beta}{R_1'}d\tau\\&=-\nabla\varphi_1-\nabla_2\varphi_2,\end{aligned} \tag{2}$$

[1] J. N. Goodier, "On the integration of the thermoelastic equations," Phil. Mag. 7 (No. 23), 1017–1032 (1937).
[2] N. O. Myklestad, "Two problems of thermal stress in the infinite solid," J. App. Mech. 9, A-136-143 (1942).
[3] R. D. Mindlin and D. H. Cheng, "Nuclei of strain in the semi-infinite solid," J. App. Phys. 21, 926 (1950).
[4] R. D. Mindlin, "Force at a point in the interior of a semi-infinite solid," Physics 7, 195–202 (1936).

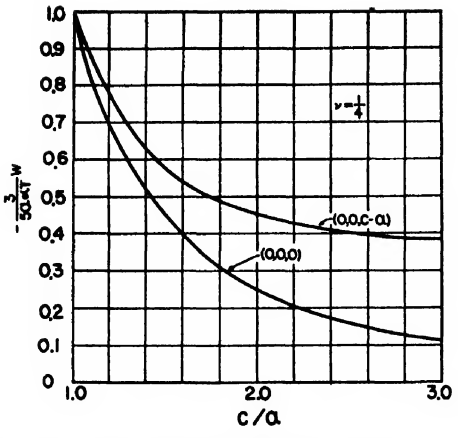


FIG. 1. Displacement normal to surface at points (0, 0, 0) and (0, 0, $c-a$).

where $d\tau=d\xi d\eta d\zeta$ and V_2 is the image of V_1 in the plane $z=0$. The volume integrals pass through ∇ and ∇_2 since the latter are x, y, z operators. The first integral is the potential (φ_1) of a distribution of matter of density β in V_1 as in Goodier's theory. The second integral (φ_2) is simply the reflection transformation of φ_1 in the plane $z=0$, so that the integration need not be repeated. For any distribution of β whose potential is known, the calculation of displacement in the semi-infinite body thus reduces to the simple process of applying the operators ∇ and ∇_2. Although φ_1 and φ_2 are mutual images, the corresponding displacements are not. This is due to the difference between ∇ and ∇_2.

III. APPLICATION TO SPHERICAL INCLUSION

Let V_1 be a sphere of radius a with center at (0, 0, c) where $a\leqslant c$. The radii to a field point (x, y, z) from (0, 0, c) and from its image (0, 0, $-c$) are denoted by $R_1=[x^2+y^2+(z-c)^2]^{\frac{1}{2}}$ and $R_2=[x^2+y^2+(z+c)^2]^{\frac{1}{2}}$ respectively. Let α be the excess of the coefficient of thermal expansion of the material within V_1 over that of the surrounding medium and let T be a uniform

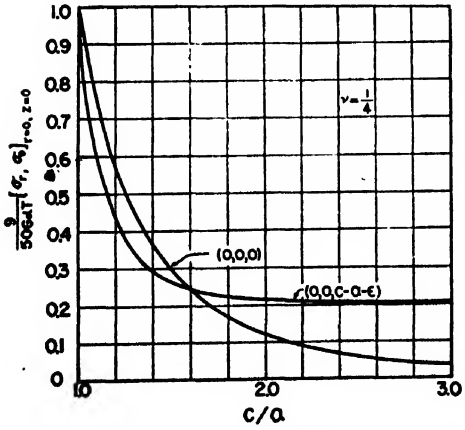


FIG. 2. σ_r and σ_θ at points (0, 0, 0), (0, 0, $c-a-\epsilon$).

temperature rise. Then β is given by (1) in $R_1<a$ and is zero in $z\geqslant 0$, $R_1>a$. The elastic constants of the two regions are assumed to be the same.

From the well-known results for the potential of a homogeneous sphere[5] we have, for exterior points ($z\geqslant 0$, $R_1>a$)

$$\varphi_1=4\pi a^3\beta/3R_1, \quad \varphi_2=4\pi a^3\beta/3R_2$$

and, for interior points ($R_1<a$)

$$\varphi_1=2\pi\beta(3a^2-R_1^2)/3, \quad \varphi_2=4\pi a^3\beta/3R_2.$$

Substituting in (2) and performing the operations ∇ and ∇_2, we have, for exterior points,

$$\mathbf{u}_e=\frac{a^3\beta}{3}\left\{\frac{\mathbf{R}_1}{R_1^3}+\frac{(3-4\nu)\mathbf{R}_2}{R_2^3}-\frac{6z(z+c)\mathbf{R}_2}{R_2^5}-\frac{2\mathbf{k}}{R_2^3}[(3-4\nu)(z+c)-z]\right\},$$

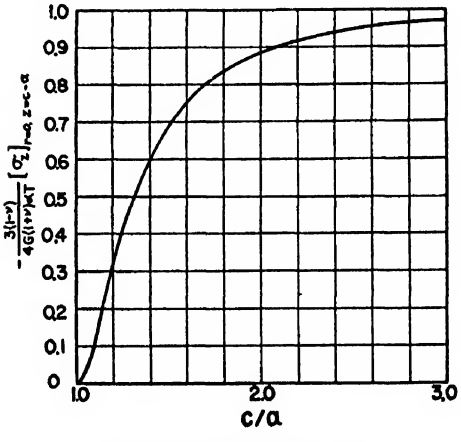


FIG. 3. σ_z at point (0, 0, $c-a$).

and, for interior points,

$$\mathbf{u}_i=\mathbf{u}_e+\frac{a^3\beta\mathbf{R}_1}{3}\left(\frac{1}{a^3}-\frac{1}{R_1^3}\right),$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are the position vectors from (0, 0, c) and (0, 0, $-c$), respectively. It may be observed that, for exterior points, the displacement is the same as that for a single center of dilatation at the center of the sphere.

The stress is calculated from the displacement by means of the usual thermoelastic relation. In dyadic notation[6] this is

$$\Psi=\lambda\mathbf{I}[\nabla\cdot\mathbf{u}-\beta(1-\nu)/\nu]+G(\nabla\mathbf{u}+\mathbf{u}\nabla)$$

where Ψ is the stress dyadic, $\mathbf{I}$ is the idemfactor and

[5] W. D. MacMillan, *Theory of the Potential* (McGraw-Hill Book Company, Inc., New York, 1930), p. 40.
[6] C. E. Weatherburn, *Advanced Vector Analysis* (G. Bell and Sons, London, 1928).

$\lambda=2G\nu/(1-2\nu)$. In performing the calculations for interior points, β is given by (1) while for exterior points $\beta=0$. The cylindrical components are, for exterior points,

$$\sigma_r=\sigma_\theta-2G\beta a^3r^2\left[\frac{1}{R_1^5}+\frac{3-4\nu}{R_2^5}-\frac{10z(z+c)}{R_2^7}\right],$$

$$\sigma_\theta=\frac{2G\beta a^3}{3}\left[\frac{1}{R_1^3}+\frac{3-8\nu}{R_2^3}-\frac{6z(z+c)}{R_2^5}+\frac{12\nu(z+c)^2}{R_2^5}\right],$$

$$\sigma_z=\frac{2G\beta a^3}{3}\left[\frac{1}{R_1^3}-\frac{1}{R_2^3}-\frac{18z(z+c)}{R_2^5}+\frac{3(z+c)^2}{R_2^5}-\frac{3(z-c)^2}{R_1^5}+\frac{30z(z+c)^3}{R_2^7}\right],$$

$$\tau_{rz}=-2G\beta a^3r\left[\frac{z-c}{R_1^5}+\frac{3z+c}{R_2^5}-\frac{10z(z+c)^2}{R_2^7}\right],$$

$\tau_{\theta r}=\tau_{z\theta}=0$.

For interior points the formulas are the same except that in each of the normal components all the terms containing R_1 are replaced by $-2/a^3$, while the term containing R_1 in the shear component vanishes.

The magnitudes of the components of stress and displacement at the point on the sphere nearest the free surface and at the point on the free surface nearest the sphere are plotted in the Figs. 1, 2, and 3 as functions of the depth of the center of the sphere below the surface. It is interesting to observe that for an expanding sphere close to the surface the tensile stress in the surrounding medium can become $4(1+\nu)$ times as great as for the sphere in an infinite medium.

The Generalized Response of Linear Systems for Arbitrary Initial Conditions

JOHN R. MOORE
Aerophysics and Atomic Energy Division, North American Aviation, Inc. and Department of Engineering, University of California, Los Angeles, California

(Received May 1, 1950)

A method is presented for the solution of ordinary linear differential equations with constant coefficients. This method stays wholly in the time domain and obtains a general solution in a compact form including the effects of initial conditions. The differential equation as treated involves a driving function made up of the sum of terms involving the system input and its derivatives. The general solution is also given in terms of weighting functions operating on the input (in the driving function) and including outside the weighted integral, terms involving initial conditions of the input as well as the dependent variable.

INTRODUCTION

THIS paper presents a short method of computing the response of single input-single output linear systems with constant parameters to a large variety of driving functions. The result of this method is a general formula for the response expressed in terms of integrals of the input, the classical exponential transient terms and initial values of both the input and response functions together with their derivatives. The mathematical development is primarily in terms of functions in the time domain. The method and formulas are equally applicable to any linear differential equations with constant coefficients whether it be thought of in terms of a time dependent "input" and "response" or merely as a problem in mathematics.

SOLUTION OF THE DIFFERENTIAL EQUATION

The present paper takes the differential equation

$$G(p)\theta=KH(p)i \qquad (1)$$

[where θ is the dependent variable, i the input (function of variable t) $p\equiv d/dt$ and $G(p)$ and $H(p)$ are polynomials in p with constant coefficients] and obtains the general solution for θ, including the effect of initial conditions on θ and its derivatives.

The solutions may be made applicable to the case of multiple roots (two or more of the R_j equal) by standard limiting procedures.

The present method takes its cue from the classical method of successive substitution.* It was originally derived by the writer following a realization that the order of successive substitution is completely arbitrary and may be rearranged at will.

Starting with $G(p)$ in its factored form, n similar quantities, Θ_j are defined such that,

$$(p-R_j)\Theta_j\equiv G(p)\theta=KH(p)i. \qquad (2)$$

This is, apparently, equivalent to

$$\Theta_j=\frac{G(p)\theta}{p-R_j}=\frac{\prod_{i=1}^{n}(p-R_i)\theta}{p-R_j}=\frac{-\partial G(p)\theta}{\partial R_j}. \qquad (3)$$

* This is similar to the "Factorization Method" described in *Elementary Differential Equations*, by T. C. Fry (D. Van Nostrand Company, Inc., New York, 1929), seventh printing, pp. 182–186.

Solving (2) for Θ_j, n equations are obtained of the form

$$\Theta_j(t)=\Theta_j(0)e^{R_jt}+K\int_0^t e^{R_j(t-\tau)}H(p)i(\tau)d\tau. \quad (4)$$

Here

$$\Theta_j(0)=\frac{G(p)}{p-R_j}\theta\bigg|_{t=0}=\frac{-\partial G(p)}{\partial R_j}\theta\bigg|_{t=0}\equiv\frac{-\partial G(p_0)}{\partial R_j}\theta, \quad (5)$$

where

$$p_0{}^k\theta\equiv(d^k\theta/dt^k)\,|_{t=0}.$$

Since, by (2), each Θ_j contains θ and its derivatives through $p^{n-1}\theta$, it is possible to solve for θ from the n equations typified by (4) obtaining a linear combination of the various Θ_j. Although this could be done by the method of determinants, it is much simpler to proceed as follows.

Assume the linear combination is

$$\theta(t)=\sum_{v=1}^{n}k_v\Theta_v(t). \quad (6)$$

From Eq. (3),

$$\Theta_v(t)=\prod_{r=1}^{n\ (v)}(p-R_r)\theta(t), \quad (7)$$

where the superscript on the product symbol denotes omission of the term for which $v=r$. Substitution of (7) into (6) gives:

$$\theta(t)\equiv\sum_{v=1}^{n}k_v\prod_{r=1}^{n\ (v)}(p-R_r)\theta(t). \quad (8)$$

Equation (8) is an identity which must hold not only for all t but also for all functions $\theta(t)$. In particular, let $\theta(t)=e^{R_jt}$:

$$e^{R_jt}\equiv\sum_{v=1}^{n}k_v\prod_{r=1}^{n\ (v)}(R_j-R_r)e^{R_jt}. \quad (9)$$

All of the products in Eq. (9) contain (R_j-R_j) as a factor except when $v=j$. Hence:

$$k_j=\frac{1}{\prod_{r=1}^{n\ (j)}(R_j-R_r)}=\frac{1}{-\lim_{p\to R_j}[\partial G(p)/\partial R_j]}$$

$$\equiv\frac{1}{-[\partial G(R_j)/\partial R_j]}. \quad (10)$$

Substituting (10) and (4) into (6) gives the general formula:

$$\theta=\sum_{j=1}^{n}\left\{\frac{e^{R_jt}\dfrac{\partial G(p_0)}{\partial R_j}\theta-K\displaystyle\int_0^t e^{R_j(t-\tau)}H(p)i(\tau)d\tau}{[\partial G(R_j)/\partial R_j]}\right\}. \quad (11)$$

INTRODUCTION OF INPUT INITIAL CONDITIONS

Equation (11) contains $H(p)i$. This is often not as convenient to use as another form containing $H(R_j)i$ with initial conditions on i and its derivatives. To derive this form from (11), $H(p)$ is written as a summation, giving

$$\sum_{j=1}^{j=n}k_j\int_0^t e^{R_j(t-\tau)}H(p)i(\tau)d\tau$$

$$=\sum_{j=1}^{j=n}\sum_{\alpha=0}^{\alpha=m}\int_0^t k_je^{R_j(t-\tau)}h_\alpha p^\alpha i(\tau)d\tau. \quad (12)$$

Integrating each term by parts, gives elements of the form

$$\int_0^t e^{R_j(t-\tau)}p^\alpha i(\tau)d\tau$$

$$=(p^{\alpha-1}+R_jp^{\alpha-2}+\cdots+R_j{}^{\alpha-2}p+R_j{}^{\alpha-1})i(t)$$

$$-e^{R_jt}(p_0{}^{\alpha-1}+R_jp_0{}^{\alpha-2}+\cdots+R_j{}^{\alpha-2}p_0+R_j{}^{\alpha-1}p_0{}^0)i$$

$$+R_j{}^\alpha\int_0^t e^{R_j(t-\tau)}i(\tau)d\tau. \quad (13)$$

The form of (13) may be simplified by using the relation

$$(x-R_j)[x^{\alpha-1}+R_jx^{\alpha-2}$$

$$+\cdots+R_j{}^{\alpha-2}x+R_j{}^{\alpha-1}]=x^\alpha-R_j{}^\alpha \quad (14)$$

so that, symbolically

$$(p^{\alpha-1}+\cdots+R_j{}^{\alpha-1})i(t)\equiv[(p^\alpha-R_j{}^\alpha)/(p-R_j)]i(t) \quad (15)$$

and

$$e^{R_jt}(p_0{}^{\alpha-1}+\cdots+R_j{}^{\alpha-1}p_0{}^0)i$$

$$\equiv e^{R_jt}[(p_0{}^\alpha-R_j{}^\alpha p_0{}^0)/(p_0-R_j)]i. \quad (16)$$

Substitution of these equations into (12) gives

$$\sum_{j=1}^{j=n}k_j\int_0^t e^{R_j(t-\tau)}H(p)i(\tau)d\tau$$

$$=\sum_{j=1}^{j=n}k_jH(R_j)\int_0^t e^{R_j(t-\tau)}i(\tau)d\tau$$

$$+\sum_{j=1}^{j=n}k_j\sum_{\alpha=0}^{\alpha=m}h\alpha\left[\frac{p^\alpha-R_j{}^\alpha}{p-R_j}-e^{R_jt}\frac{(p_0{}^\alpha-R_j{}^\alpha p_0{}^0)}{p_0-R_j}\right]i$$

$$=\sum_{j=1}^{j=n}k_j\left\{H(R_j)\int_0^t e^{R_j(t-\tau)}i(\tau)d\tau\right.$$

$$\left.-e^{R_jt}\left[\frac{H(p_0)-H(R_jp_0{}^0)}{p_0-R_j}\right]i+\left[\frac{H(p)-H(R_j)}{p-R_j}\right]i\right\}. \quad (17)$$

This may be simplified by noting that the relations between the k_j and R_j make

$$\sum_{1}^{n} k_j R_j^{\alpha} = 0, \quad (0 \leq \alpha \leq n-2). \tag{18}$$

Thus, (17) becomes

$$\sum_{j=1}^{j=n} k_j \int_0^t e^{R_j(t-\tau)} H(p) i(\tau) d\tau$$

$$= \sum_{j=1}^{j=n} k_j \left\{ H(R_j) \int_0^t e^{R_j(t-\tau)} i(\tau) d\tau \right.$$

$$- e^{R_j t} \left[\frac{H(p_0) - H(R_j p_0^0)}{p_0 - R_j} \right] i$$

$$\left. + R_j^{n-1} \sum_{\beta=n}^{\beta=m} h_\beta \frac{(p^{\beta-n+1} - R_j^{\beta-n+1}) i}{p - R_j} \right\} \tag{19}$$

which makes the final solution for θ

$$\theta = -\sum_{1}^{n} \left[\frac{\partial G(R_j)}{\partial R_j} \right]^{-1}$$

$$\times \left\{ \frac{\{G(p_0)\theta - K[H(p_0) - H(R_j p_0^0)]i\}}{p_0 - R_j} \right\} e^{R_j t}$$

$$+ KH(R_j) \int_0^t e^{R_j(t-\tau)} i(\tau) d\tau$$

$$\left. + R_j^{n-1} \sum_{\beta=n}^{\beta=m} h_\beta \frac{(p^{\beta-n+1} - R_j^{\beta-n+1}) i}{p - R_j} \right\}, \tag{20}$$

where the last summation exists only if $H(p)$ is of the same order as, or higher order than $G(p)$.

CONCLUSION

The foregoing development attempts to carry out the general solution of the differential equation of an idealized linearized system including the often neglected effect of initial conditions both on the input and the response. The end formula is only slightly more difficult to remember than that of the Heaviside expansion theorem, yet covers a more general case. The method of solution is much simpler than the method of successive substitutions and yields more usable results for high order equations than the more usual type of Laplace transform solution.

BIBLIOGRAPHY

(1) G. Doetsch, *Theorie und Anwendung der Laplace-Transformation* (Dover Publications, New York, 1943, U. S. Alien Property Custodian), pp. 321–328.

(2) T. C. Fry, *Elementary Differential Equations* (D. Van Nostrand Company, Inc., New York, 1929), pp. 182–186.

(3) M. Gardner and J. L. Barnes, *Transients in Linear Systems* (John Wiley and Sons, Inc., New York, 1945).

(4) K. Wagner, *Operatorenrechnung* (Edwards Brothers, Inc., Michigan, 1944, U. S. Alien Property Custodian), pp. 94–98.

(5) Other standard books on differential equations.

Letters to the Editor

The Thermal Expansion Coefficient and the Melting Point of Cubic Elements

M. E. Straumanis
University of Missouri, School of Mines and Metallurgy, Department of Metallurgy, Rolla, Missouri
April 10, 1950

IN the Fig. 1 the linear expansion coefficient, as defined by $\alpha=(1dl)/(ldT)$, of cubic elements is plotted against the melting point T_F of the same elements.[1] The curve shows clearly that generally the linear expansion coefficient of cubic elements decreases with increasing melting point. There are no exceptions from this rule, only deviations from the curve. These deviations might be caused (1) by some differences in the structure of the cubic elements, or (2) by inaccurate determinations of the expansion coefficients at room temperature. Further work will show the nature of these deviations. The curve presented seems not to be a simple one, although the melting points of the 5 alkali elements fit well to the empirical equation $\alpha\sqrt[3]{T_F}=\text{const}$, T_F being expressed in °C.

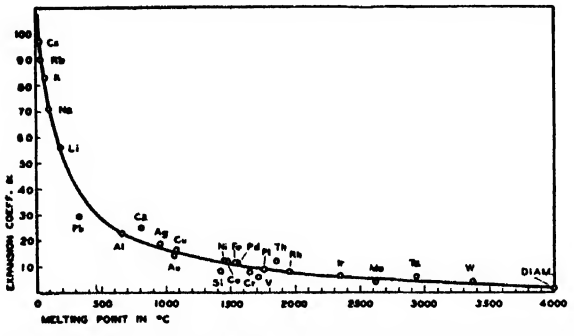


FIG. 1. Linear expansion coefficient $\alpha\times10^6$ of cubic elements at room temperature *versus* melting point of the same elements.

It might be that plotting the cubical expansion coefficient in the same manner as already shown, all at room temperature solid elements would be located on or near the curve.

No relations between melting point and expansion coefficient could be found in the literature, except some indications made by Grüneisen[2] and Lindemann.[3]

[1] The coefficients and melting points in Fig. 1 were obtained from the International Critical Tables, Vol. I, 103, from Hodgman's *Handbook of Chemistry and Physics* and from Landoldt-Börnstein, *Physikalischechemische Tabellen*.
[2] E. Grüneisen, Ann. d. Physik 39, 297 (1912).
[3] F. A. Lindemann, Physik. Zeits. 11, 609 (1910).

The Tracing and Interpretation of Asymmetrical Hysteresis Loops*

G. M. Ettinger**
Department of Electrical Engineering, New York University, New York, New York
May 3, 1950

IN the course of development of a hysteresis loop tracer with cathode-ray tube presentation (to be described elsewhere), the requirement arose to display the asymmetrical hysteresis loops[1] which are exhibited by magnetic materials subjected simultaneously to a.c. and d.c. magnetizing forces. Figures 1a to 1e show a series of loops which were obtained for "permalloy C" with increasing d.c. polarization. The polarizing current was produced by connecting in series with the magnetizing winding on the sample a metal rectifier shunted by a variable resistance.

To trace the curve of incremental permeability *vs.* d.c. polarizing field, the a.c. sensitivity of the (direct coupled) horizontal deflection amplifier in the hysteresis loop tracer was reduced to zero. The trace on the cathode-ray tube therefore changed to a straight line whose length was proportional to the peak flux density in the sample, while its displacement from the axis of B (shown dotted in Fig. 1) gave a measure of the d.c. magnetizing force. Such lines are superposed on their corresponding hysteresis loops in Fig. 1, while Fig. 2 shows a number of these lines photographed on the same film. Their envelope is the curve of $B_{a.c.}$ *vs.*

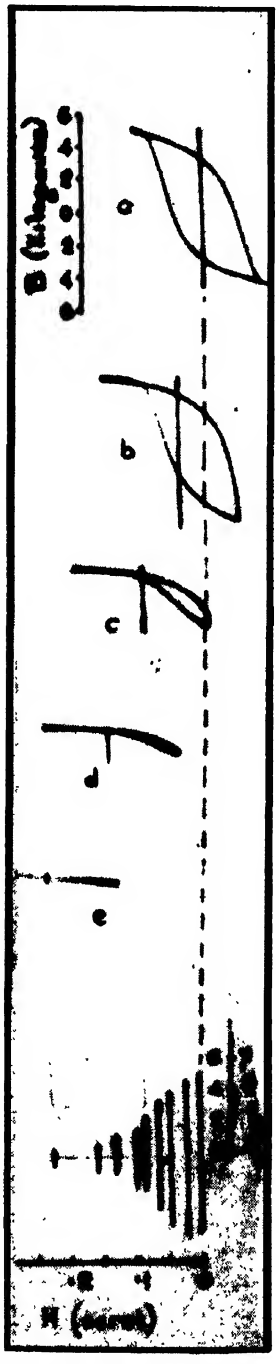


FIG. 1. Asymmetrical hysteresis loops.

FIG. 2. A.c. permeability *vs.* polarizing field for "permalloy C."

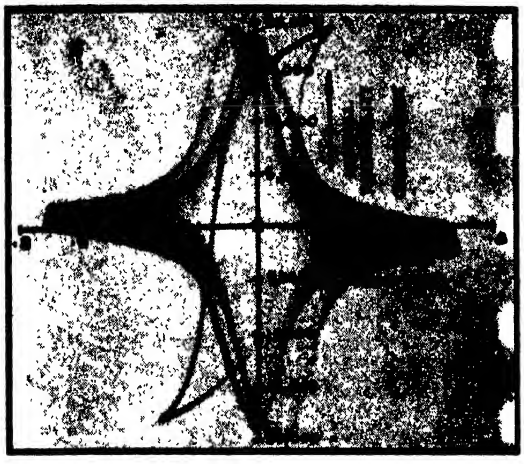


FIG. 3. Butterfly curve for "permalloy C."

$H_{d\,c}$. The peak value of the alternating component of magnetizing force was kept constant during the test, so that Fig. 2 represents, to a different scale, a portion of the $\mu_{a\,c}/H_{d\,c}$ or "butterfly" curve which is of importance in magnetic amplifier work. Other workers[2,3] have employed bridge or potentiometer methods to obtain this curve.

It is possible to observe the complete butterfly curve (from positive to negative saturation) on the screen of a long persistence cathode-ray tube if the potentiometer controlling the d.c. polarization is slowly driven to and fro by a reversing motor or a cam Figure 3 is a photograph of the "butterfly curve" for "permalloy C," obtained by this method. A hysteresis loop is superposed for calibration purposes.

[1] L. F. Borg, J. Inst. Elec. Eng. (London) 96 (Part II), 317 (1949).
[2] G. W. Elmen, Nat. Bur. Stand. Tech. J. 15, 124 (1936).
[3] B. Fertchak, Rev. Gen. Elec. (Paris) 54, 79 (1945).
* This work was done at Standard Telecommunication Laboratories, Ltd., London, England.
** Department of Electrical Engineering, New York University.

Mechano Chemistry of the Dispersion of Mercury in Liquids in an Ultrasonic Field

EVELYN C. MARBOE AND W. A. WEYL
The Pennsylvania State College, School of Mineral Industries, State College, Pennsylvania
May 5, 1950

IN a previous publication[1] on the mechano-chemical properties of water it was pointed out that the strength of the binding forces between water and a metal may play a role in certain types of cavitation corrosion. It is the object of this report to describe some simple experiments which illustrate the role which the binding forces between metallic mercury and a liquid play on the dispersion of this metal in the liquid. For our problem, the use of liquid mercury offers the advantage that there can be no work-hardening of the metal, a factor which complicates the corrosion through cavitation for solid metals. Some parallel experiments were carried out with metallic gallium in order to make certain that the phenomena observed were not characteristic for mercury alone. All experiments were carried out with a quartz crystal type of high frequency generator. (Ultrason Model U-300 of the Televiso Products Company, Chicago, Illinois.)

Experiment No. 1. For a certain intensity of the ultrasonic field (0.2 kv, 20 ma at 25°C) the system, mercury-paraffin oil, was found to remain unaffected. Addition of stearic acid to the paraffin oil, however, caused dispersion of the metal. This experiment which confirms previous work[2] indicates clearly the importance of the forces acting between the metal and the hydrocarbon. A molecular film of Hg-stearate links together the incompatible media and, thus, makes dispersion possible and also stabilizes the emulsion by preventing the droplets from coalescing.

Experiment No. 2. The intensity of the ultrasonic field was adjusted (1 kv, 130 ma at 25°C) so that mercury in distilled water produced a slight turbidity after an exposure of five minutes. Under these conditions the water remained clear if it had been deaired by boiling. The effect of the air on the dispersion of the metal in water is explained as follows.

Oxidation of the mercury produces Hg^{+} and Hg^{2+}, both cations having a high polarizability. As has been pointed out in a previous report[3] ions of this type are adsorbed at a metal surface in a fashion which may be described as:

Metal—metallic bond—polarized cation—ionic bond—water.

By changing the metallic character of the metal surface into one resembling that of an ionic substance, the strong binding forces are provided between metal and water which are essential for the dispersion.

Experiment No. 3. The same experiment was repeated with distilled and deaired water. No turbidity was observed after five minutes exposure. Now a few drops of a saturated $HgCl_2$ solution were added in order to provide Hg^{2+} ions. Immediately a deep gray dispersion of Hg droplets in the diluted $HgCl_2$ solution began to form.

The connecting link between metal and water is the Hg^{2+} ion which becomes strongly polarized in the asymmetrical force field of the metal-water interface. Addition of KCl solution to the Hg^{2+} containing water changes the Hg^{2+} into a complex anion of the type $(HgCl_4)^{2-}$ and, as a result, mercury is no longer easily dispersed. Only the cation Hg^{2+}, but not the anion $(HgCl_4)^{2-}$, has the metallophilic properties essential for producing dispersion and stabilizing the emulsion. This experiment has its analog in the effect of KCl upon the solubility of metallic lead in fused lead chloride.

R. Lorenz[4] determined the influence of alkali chloride on the solubility of metallic lead in fused $PbCl_2$ at 610°C (Table I).

TABLE I. Solubility of metallic lead in the system $PbCl_2$—KCl.

$PbCl_2$	KCl	Solubility X10,000
10 mole	0 mole	3.74%
9 mole	1 mole	2.27%
8 mole	2 mole	1.51%
7 mole	3 mole	0.64%
6 mole	4 mole	0.06%

The solubility of the metal in the fused salt is the result of the presence of Pb^{2+} ions which, in the asymmetrical force field of the Pb—$PbCl_2$ interface, are strongly deformed and thus are able to link the metal atoms with the ionic medium. Addition of KCl to the melt changes the $PbCl_2$ into K_2PbCl_4. The anion $(PbCl_4)^{2-}$ cannot act as a metallophilic group.

Experiment No. 4. Liquids containing polar molecules exert image forces upon the mercury and, as a result, increase the dispersability of the mercury in a liquid. Under conditions which did not lead to a dispersion of Hg metal in the non-polar carbon tetrachloride, the metal could be dispersed in the strongly polar nitrobenzene.

Summary. Dispersion of Hg in a liquid under the influence of ultrasonic vibrations requires the presence of strong forces between the metal and the medium. Strengthening the forces between the two incompatible media enhances dispersion and stabilizes the emulsion. This can be achieved in three ways:

1. Compound formation between the metal and a part of the liquid, e.g., addition of stearic acid to paraffin oil leading to a molecular film of the metal stearate.

2. Image forces induced in the metal surface by molecules with a permanent dipole, e.g., nitrobenzene.

3. Metallophilic groups such as strongly polarizable cations, e.g., Hg^{2+}. The mechanism of their action is known from several phenomena involving adhesion[5] (silver to glass by means of tin ions) and solution[6] (gold ruby glass requires the presence of metallophilic ions, such as Sn^{4+}, Bi^{3+}, Pb^{2+}).

[1] W. A. Weyl and E. C. Marboe, Research 2, 19 (1949).
[2] L. N. Solov'eva, Colloid. J. (U.S.S.R.) 5, 289–97 (1939).
[3] Enright, Marboe, and Weyl, ONR Tech. Report No. 1, Contract No. N6 onr 269, Task Order 8 NR 032-265, 11 pages (October, 1948).
[4] R. Lorenz and W. Eitel, *Pyrosole* (Academ. Verlags-ges., Leipzig, 1926), in particular pp. 54 and 55.
[5] W. A. Weyl, The Glass Industry 26 (12), 557 (1945).
[6] W. A. Weyl, J. Soc. Glass Tech., "Colors Produced by Metals. Part IV." 29, 302–307 (1945).

The Barium Oxide on Tungsten Cathode Interface*

EUGENE B. HENSLEY AND JOHN H. AFFLECK
University of Missouri, Columbia, Missouri
May 17, 1950

IT is now recognized that the electrical behavior of oxide coated cathodes may be influenced either directly or indirectly by the presence of an interface compound formed by the reaction of the oxide coating with the base metal.[1] Tungsten and tungsten nickel alloys have been used as the base metal in experimental tubes and have found some commercial application in certain subminiature tubes. We have identified the interface compound for these cathode systems as being composed principally of barium tungstite, $BaWO_3$, when the coating is pure BaO. The corresponding tungstites are found when SrO or the solid solution (BaSr)O are used for the coating. An extensive search of the literature has failed to reveal any mention of the existence of these compounds; consequently particular care has been taken in their identification. Because of the small amounts of the interface compounds formed, x-ray diffraction methods of analysis have been used.

Cathodes were prepared by spraying $BaCO_3$ with an organic binder onto 10 mil tungsten wires which had been cleaned by electrolyzing in a KOH bath. The cathodes were then mounted in vacuum tubes and exhausted to a final pressure of 5×10^{-8} mm of Hg. The $BaCO_3$ was then reduced to the oxide by passing heater current through the tungsten wires and the tubes were sealed off with the pressure again at 5×10^{-8} mm. The tubes were then placed on life test for various periods of time and at various cathode temperatures. No emission current was drawn from any of the experimental cathodes included in this report.

After aging for a definite period of time the tubes were broken open, the oxide coatings were removed mechanically and x-ray diffraction patterns were made of the exposed interface. In order to be sure the interface compound was stable in air, the first samples were opened in a dry-box and then coated with a protective coating of waterproof wax. No change in the interface pattern was noticed between the cathodes treated in this manner and those which were not. The x-ray powder diffraction patterns were made with a camera of 7.16 cm radius and using Cu Kα-radiation filtered through 0.0007 in. nickel.

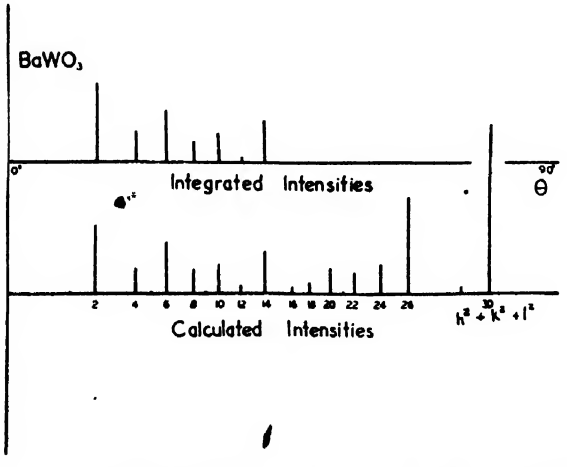


FIG. 1. The integrated intensities and Bragg angles of the x-ray powder diffraction lines of $BaWO_3$ compared with their calculated values.

The x-ray diffraction lines showed the interface compound to have a cubic structure with a lattice spacing of 4.30 *kX*. In addition to these and the tungsten lines, a few other lines were sometimes present which were probably due to tungsten oxides but they were always too weak to permit identification. Assuming the interface compound to have a perovskite ($CaTiO_3$) structure with a formula $BaWO_3$, the intensities were calculated and are shown in Fig. 1 together with the integrated intensities obtained by microphotometering the x-ray diffraction films. The calculation of these intensities involved the use of absorption factors for coated cylinders.[2] The lines beyond a Bragg angle of 45° were too broad to be measured. This broadening was probably due to the small crystal sizes and also to the fact that most perovskite type crystals are pseudo-cubic, causing the lines in the back reflections to have a multiple structure.

Cathodes consisting of SrO on W were also prepared and the x-ray analysis indicated the presence of $SrWO_3$ with a lattice constant of about 4.12 *kX*.** These lines were much weaker than those obtained from the barium compound, indicating the BaO reacts more easily with the W than does the SrO. Also prepared were cathodes with a coating of equal molar solid solution (BaSr)O on W. The lattice constant of the resultant $(BaSr)WO_3$ was 4.24 *kX*, indicating greater than an equal molar fraction of barium. Similar results to the above have been obtained when the cathode was formed on a 3 mm diameter nickel sleeve containing 4.7 percent tungsten impurity. Commercial hearing aid tubes[3] using a 0.5 mil tungsten wire base for the oxide cathode have been examined and the tungstite interface was observed to be present both in new tubes and in tubes which had been operated for 500 hours.

In order to determine whether the reaction requires the presence of any other material than the BaO and W, a special tube was constructed in which the BaO was deposited onto clean W by evaporation. $BaCO_3$ was sprayed onto a small spiral of pure platinum wire[4] which was then mounted in the vacuum tube with the tungsten wire passing along the axis of the spiral. First the $BaCO_3$ was reduced to BaO by passing current through the Pt wire. The W wire was then cleaned by raising its temperature to 1800°C. BaO was evaporated onto the cold W by raising the temperature of the Pt to about 1400°C until no oxide remained visible on the Pt. The W wire with its coating of BaO was heated to 850°C for 24 hours after which the cathode was removed and placed in the x-ray diffraction camera. X-ray diffraction patterns showed that $BaWO_3$ had been formed as in the previous tubes.

Several attempts were made to prepare $BaWO_3$ in larger quantities by heating mixtures of various tungsten and barium compounds in a vacuum furnace as for example powdered tungsten and barium carbonate. All these experiments resulted in forming the tungstate $BaWO_4$ instead of the tungstite $BaWO_3$. The lowest pressures obtainable in the vacuum furnace were of the order of 10^{-5} mm of Hg. When a mixture of powdered W and $BaCO_3$ was coated on a cathode sleeve and heated in a vacuum of about 10^{-7} mm the x-ray analysis showed a large percentage of $BaWO_3$ present and no $BaWO_4$. From these experiments it seems probable that the tungstite is formed only in a good vacuum and at higher pressures the tungstate is formed.

An investigation of chemical properties of $BaWO_3$ has not yet been carried out. However, it has been observed that although the tungstite is relatively stable when exposed to air over a period of several months, if a sample is placed in water for several minutes much of the tungstite oxidizes over into the tungstate.

A series of identical tubes were prepared and life tested at a temperature of 875°C in order to determine the variation in interface thickness with life. An x-ray method was employed for these thickness measurements.[2] The times recorded include the

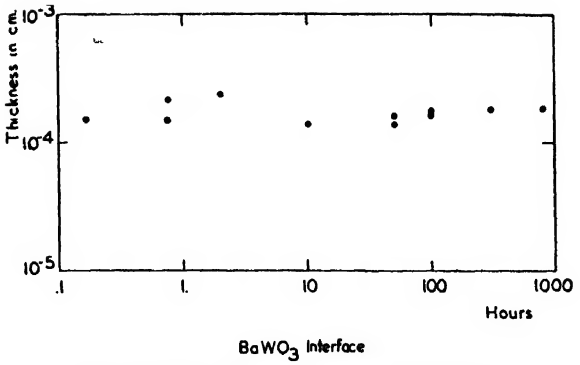


FIG. 2. Variation of interface thickness as a function of time at a temperature of 875°C.

time required to reduce the $BaCO_3$ to the oxide. Results are presented in Fig. 2 for the 15 cathodes examined. It can be seen that at this temperature the interface reaches its final thickness during the conversion process (less than 10 minutes) and remains essentially constant at 1.5×10^{-4} cm during the subsequent life of the tube. It should be pointed out that if the tungsten is present only as an impurity in another metal such as nickel, the rate of diffusion of the W may lead to an increase in thickness with life. Also drawing emission current from the cathode may well influence the interface thickness.

We wish to extend our thanks to Professor A. S. Eisenstein for many helpful discussions, and also to Mr. Harold John, who carried out the chemical preparations and to Mrs. Hope Lorch, who constructed most of the tubes.

* This work supported in part by ONR.
[1] A. S. Eisenstein, *Advances in Electronics I* (Academic Press, Inc., New York, 1948), p. 24.
[2] E. B. Hensley, Phys. Rev. 77, 744(A) (1950).
** The kX unit used is about the same as 10^{-3} cm.
[3] Type CK522AX supplied by I. E. Levy of Raytheon Manufacturing Company.
[4] Prepared by H. T. Reeve of Bell Telephone Laboratories, Inc.

An Alternative Method for the Summation of Fourier Series

I. F. MORRISON
Department of Civil and Municipal Engineering, University of Alberta, Edmonton, Alberta, Canada
May 25, 1950

AS an alternative to the interesting method for the summation of certain types of Fourier Series presented in the April issue by L. A. Pipes, a more direct approach can be made by using summation formulas for trigonometric series in terms of the Bernoulli and Euler polynomials. There are four of these formulas, which are, perhaps, not quite as well known as they might be, for they are generally quite useful, especially in such cases as are referred to in this paper.

One of them, for example, is,

$$\bar{B}_{2\nu+1}(x)=(-1)^{\nu+1}\frac{2(2\nu+1)!}{(2\pi)^{2\nu+1}}\sum_{n=1}^{\infty}\frac{\sin 2\pi nx}{n^{2\nu+1}}$$

in which $\bar{B}_{2\nu+1}(x)$ is the Bernoulli polynomial in the interval $0\leqq x<1$. The series is absolutely and uniformly convergent for values of $\nu>0$, but not for $\nu=0$ in the neighborhood of the points $x=\pm0, \pm1, \pm2, \cdots$. To illustrate, consider Example II of the paper. If one puts $\nu=0$ in the expression above, there results

$$-\bar{B}_1(x)=+\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{\sin 2\pi nx}{n}$$

which is in the same form as

$$F_2(t)=\frac{2}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi t}{s}\right),$$

Eq. (4.7) of the paper with $2x=t/s$. $-\bar{B}_1(x)=\frac{1}{2}-x$ is a periodic function with period 1. The graph starts with a value $+\frac{1}{2}$ at $x=0$ and becomes discontinuous at $x=+1$ where it experiences a discontinuity $+1$, after which the cycle is repeated. Thus, the graph, with ordinates multiplied by 2, is the same as that of Fig. 2 of the paper.

The other examples in the paper can be dealt with in the same way.

Creep-Time Law for Zinc Crystals

E. P. T. TYNDALL
Department of Physics, State University of Iowa, Iowa City, Iowa
May 25, 1950

THE creep of suitably oriented zinc single crystals has been found to follow a very simple empirical law: $S=At^M$, in which S is the strain (non-elastic) occurring in time t, counted as zero at the instant of applying the load, and A and M are constants. The crystals have been tested under constant load and therefore constant stress since the reduction in area is negligible. Series of runs have been made in which the crystal was first loaded to just beyond the linear (elastic, or Hooke's law) region and then allowed to creep. The load was then increased by a small increment and a second run made; a further increase in load produced a third run and so on for five or six runs. The constant M, which is the slope of the straight line obtained for each run when plotted as log strain against log time, is about 0.5, varying somewhat from this mean value in different runs. The values of the constant A in any one series of runs lie on straight lines when plotted as log$A$ against the stress. The positions of the lines on such a plot depend on the history of the specimen and the purity of the zinc. Their slopes are nearly the same. The simple law has been found only with complete consistency for crystals in which one of the translation directions lies very close (within about 2°) to the direction of the projection of the length of the specimen into the basal (slip) plane, or, otherwise stated, the angle between slip plane and length of specimen is the same as the angle between the length and the nearest translation direction. The law appears to hold quite accurately for series of runs extending from strains of 10^{-7}, about the least observable strain, to strains somewhat above 10^{-3}. There is no indication of a separation into the two processes generally heretofore recognized, transient and steady-state creep.

For crystals with the translation direction 5° or more out of line with the projection of the length the behavior was definitely different. Some followed the pattern of transient plus slow creep found previously[1] while others had occasional large jumps, or (in one case) stepwise creep in which rapid creep followed slow creep at regular intervals. Crystals were also found in which, after a time of rapid creep, further strain followed the creep law above but with the constant M quite small, about 0.1 or 0.2. This appears to be typical of crystals with the angle mentioned above between approximately 15° and 30°. The simple creep law may be considered as a special case of the Andrade law proposed some years ago for creep in polycrystalline lead and copper.

[1] T. A. Read and E. P. T. Tyndall, J. App. Phys. 17, 713 (1946).

Fracture Modes in High Purity Metals

T. J. AGNOR AND M. E. SHANK
Massachusetts Institute of Technology, Cambridge, Massachusetts
May 29, 1950

IT has been commonly accepted in many quarters that only iron and ferritic steels undergo a change in impact fracture mode with decreasing temperature. In such materials fracture changes from ductile type to brittle cleavage type and the absorbed energy drops markedly. The exact temperature of such

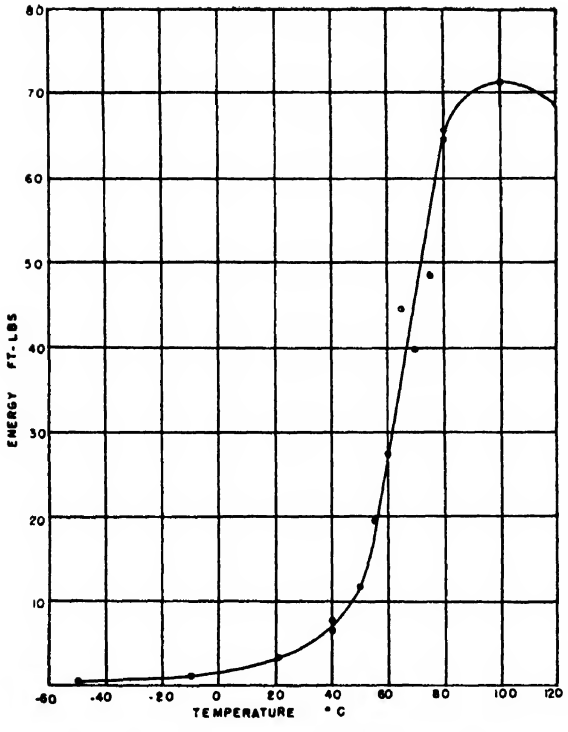


FIG. 1. Transition curve for high purity polycrystalline zinc.

transition depends upon composition and heat treatment of the material. Investigation by the authors into the behavior of high purity zinc by means of standard Charpy tests shows similar behavior of this metal. The material used was of 99.9955 percent purity, containing 0.0021 percent Pb, 0.0017 percent Cd, 0.0006 percent Fe, and 0.0001 percent Sn. It was hot-worked to an average grain diameter of 0.013 mm. The transition curve is shown in Fig. 1. The fractured surfaces, Fig. 2, have a macroscopic appearance that is superficially indistinguishable from that found in ferritic steels. Moreover, microscopic examination of the zinc fractures indicates brittle cleavage facets adjacent to large areas of ductile fracture in specimens whose gross macroscopic fracture appearance is entirely ductile. In some instances, large, deep voids have been opened in the material, apparently due to the triaxial stress conditions developed during fracture. It is evident, therefore, that transition temperatures exist for high purity metals other than iron and ferritic steels, and in fact for a metal not even of the body-centered lattice type.

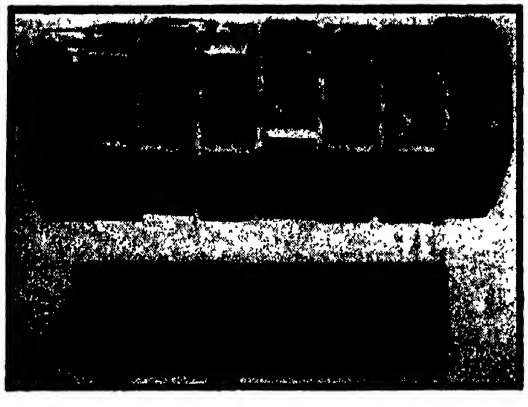


FIG. 2. Photograph of specimens showing fractured surfaces. Order of specimens, from highest to lowest temperature, is right to left on upper row, then left to right on lower row.

Similar work was carried out on high purity, vacuum-cast iron, major impurities in which were 0.063 percent O, 0.00038 percent H, 0.002 percent C. Nitrogen was not detectable. The iron had an average grain size of 0.31 mm, grain refinement being particularly difficult in such material. A very sharp transition temperature was found at 80°C. This is far above the usual transition ranges reported for ordinary Armco Iron,[1] and it does not seem probable that the difference is due entirely to the coarse grain size. On the basis of carbon and nitrogen present it would seem likely to expect a transition range in the neighborhood of −40° to −100°C. That it does not occur here apparently indicates the effect of the relatively high oxygen content. Since the work of Low and Gensamer[2] indicates that oxygen has no effect on aging or yield point in steels, the present evidence would serve to support the belief[3] that the mechanism of the transition temperature in metals is entirely unrelated to aging and inhomogeneous yielding phenomena.

[1] H. W. Gillett, "Impact Resistance and Tensile Properties of Metals at Subatmospheric Temperatures," ASTM Pamphlet, August 1941, p. 25.
[2] J. R. Low, Jr. and M. Gensamer, "Aging and the Yield Point in Steel," Trans. AIME **158**, 207 (1944).
[3] E. Orowan, "Fracture and Strength of Solids," Reports on Progress in Physics **XII**, 1949.

Intensity Correction Factors for X-Ray Spectrometer Transmission Pole Figure Determinations

SUZANNE VAN DIJKE BEATTY
Chemical Department, Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
May 25, 1950

A TABLE has been computed to expedite the reduction of intensity contours obtained with a recording x-ray spectrometer using the transmission method for pole figure determination. This method is particularly suited for thin sheet material having very small grain size and is valuable in the determination of the textures of rolled materials.

In this type of pole figure work, the counter is held stationary at a given value of 2θ corresponding to a given reflection from the sample. The sample is then rotated about the vertical axis making an angle α with the Bragg angle. For each value of α, the sample is in turn rotated about the horizontal axis making an angle β with the rolling direction. Intensity corrections are necessary since the path length in the sample varies with different values of α and θ. The correction factor at any θ is taken as unity when $\alpha=0$. It is clear that no correction is needed for different values of β since the path length in the sample remains unchanged. The final intensity depends also on the absorption of the sample and on its thickness. Fortunately we need only the product μt so that the absorption factor can be determined experimentally from the relation $e^{-\mu t}=I/I_0$. The intensity of a strong reflection from quartz, for example, is recorded both with and without the sample to be studied in front of the counter slit. The ratio of the two intensities gives the required absorption factor. The correction factor also involves the diffracting angle θ since the path length and total volume of the sample irradiated differ.

The equation giving the intensity correction factor is given by:[1]

$$\frac{I_0}{I_{\pm\alpha}}=\frac{\mu t\exp(-\mu t/\cos\theta)}{\cos\theta}\times\frac{(\cos(\theta\pm\alpha)/\cos(\theta\mp\alpha))-1}{\exp(-\mu t/\cos(\theta\pm\alpha))-\exp(-\mu t/\cos(\theta\mp\alpha))}. \quad (1)$$

This is a special case of the expression given by Smoluchowski and Turner[2] and it was found[1] that more exact values are obtained when α is negative, i.e. with clockwise rotation of the shaft of the sample holder.

The factors are given for every 5° of α from 5° to 60° inclusive. For each value of α, every 10° in 2θ is listed between 10° and 80° and every 0.1 interval of the absorption factor from 0.1 to 0.9. Families of curves are given for every 5° of α in Figs. I to XII.

The tables may serve not only as a source of correction factors, but as a guide in the preparation of the sample and data. By plotting the factors as a family of curves for each value of θ or of $e^{-\mu t}$, one can see at a glance which set of values will require the least correction. The problem of optimum settings is also discussed by L. G. Schulz[3] in a recent paper.

A limited number of these tables is available for free distribution from the Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania. Requests for the tables should be addressed directly to the author.

[1] Decker, Asp, and Harker, J. App. Phys. 19, 388-92 (1948).
[2] R. Smoluchowski and R. W. Turner, Rev. Sci. Inst. 20, 173 (1949).
[3] L. G. Schulz, J. App. Phys. 20, 1033-6 (1949).

Determination of Elastic Constants in Single Crystals with Especial Reference to Silver Chloride

DAVID L. ARENBERG
U. S. Naval Air Development Center, Johnsville, Pennsylvania
June 1, 1950

USING the ultrasonic pulse method[1] to measure the velocity of sound at 15 mc/sec. in different directions, the elastic constants of silver chloride have been determined. Two single crystals, F-16-a-2 and F-16-a-7, were oriented and ground by West and Makas of the Polaroid Corporation. After annealing carefully, the crystals were developed to produce a silver electrode on the surfaces.[2] Corrections were applied to account for the thickness of the piezoelectric transducers, but due to the large temperature and pressure coefficients of silver chloride, complete agreement between different sets was not obtained. No short time effect on the velocity was noticed when a strong light shone on the crystals. The sound absorption coefficient was low and hence not measured. The values for F-16-a-2 are believed to be better than for the set from F-16-a-7. The value 5.589 g/cc determined by Bridgman[3] from x-ray data was used for the density rather than the handbook value of 5.56.

Selected values for crystal F-16-a-2 are 0.605, 0.364, and 0.0624 times 10^{12} dynes/cm^2 for C_{11}, C_{12}, and C_{44} respectively.

Since it is not always possible to obtain orientations of crystals that allow simple calculations between the elastic constants and velocities, approximate formulas were developed and the measurements in the *hkl* directions made to check them. One can derive explicit formulas that are very good first-order approximations for all orientations and exact in certain planes and points, such as all directions in the plane normal to an 001 axis and also in the 111 direction for cubic crystals. One uses the direction cosines involving rotations about three axes with the IRE conventions given by Cady[4] of angles φ, θ, and ψ. φ and θ have their usual significance of azimuth and colatitude in spherical coordinates but ψ, which represents rotation about the Z''' axis, will be indeterminate for a pulse traveling in an infinite medium, and must be either eliminated or associated with some properties of the waves to have physical significance.

The choice of coordinates used by the experimenter is an arbitrary one which cannot influence the velocity in the medium. It is observed that if the piezoelectric transducers, particularly those generating complex waves, are mounted on the specimen the time delay between members of the series of echoes received from a single pulse will differ by increments which are integral multiples of a linear combination of the delays due to any of the three different modes having different velocities traversing the specimen once. Even with a pure mode generator operating at normal incidence, mode transfer can occur at each reflection resulting in a complex system of echoes which can, however, be analyzed into the three different fundamental modes. Changing the orientation of a shear wave transducer on the specimen can only influence the relative amplitudes of the echoes. With sufficient care in orientation, only one period between echoes will be observed.

One can therefore associate the angle ψ with the plane of polarization of the quasi-transverse waves in the crystal and by *ad hoc* reasoning apply Fermat's principle that the velocities of the waves assume stationary values so that all $\partial C_{ii}/\partial\psi=0$. The angle ψ is then uniquely determined as a function of φ and θ and will give the velocities and plane of polarization to a first-order approximation as shown below. Higher order correction terms can be added later.

One selects the expressions for C_{44}', C_{55}', and C_{33}' for operation and defining for cubic crystals:

$$A=8-\sin^2\theta(7+\cos^4\varphi) \tag{1}$$

$$B=2\cos\theta\sin^4\varphi \tag{2}$$

$$C=A-2(1-\cos^4\varphi) \tag{3}$$

$$\tan 2\psi=B/C \tag{4}$$

$$r=C_{11}-C_{12}-2C_{44}, \tag{5}$$

we can get the elastic constants along the Z''' axis:

$$C_{33}'=C_{11}-r\frac{A\sin^2\theta}{4}=C_{11}-r\alpha \tag{6}$$

$$C_{44}'=C_{44}+r\frac{\sin^2\theta}{8}(A-(B^2+C^2)^{\frac{1}{2}})=C_{44}+r\beta \tag{7}$$

$$C_{55}'=C_{44}+r\frac{\sin^2\theta}{8}(A+(B^2+C^2)^{\frac{1}{2}})=C_{44}+r\gamma. \tag{8}$$

Also for any direction in the crystal, the velocities are:

$$V_L'=(C_{33}'/\rho)^{\frac{1}{2}} \tag{9}$$

$$V_t'=(C_{44}'/\rho)^{\frac{1}{2}} \tag{10}$$

$$V_T'=(C_{55}'/\rho)^{\frac{1}{2}}. \tag{11}$$

After determining all C_{ii}' from measurements and manipulating (6), (7), and (8), one gets the inverse relations:

$$C_{11}=C_{33}'+\frac{\alpha(C_{55}'-C_{44}')}{\gamma-\beta} \tag{12}$$

$$C_{12}=C_{33}'-C_{44}'-C_{55}'-(1-2\alpha)\frac{(C_{44}'-C_{55}')}{\gamma-\beta} \tag{13}$$

$$C_{44}=\frac{\gamma C_{44}'-\beta C_{55}'}{\gamma-\beta}. \tag{14}$$

If φ and θ are known with respect to the crystal axes the three velocity measurements in one direction, such as the 110, would suffice to give all three constants except for a certain arbitrariness in the relation between γ, β, and C_{44}' and C_{55}'. For most ionic crystals, the anisotropy factor $(C_{11}-C_{12})/2C_{44}$ is >1.00 and for metallic crystals <1.00, so that this choice can be made. Errors in measurement and misorientation are least important in the 001 and 111 directions, but here $V_t'=V_T'$ and velocities must be measured in two directions.

Furthermore, with highly anisotropic crystals, one can obtain an approximate idea of the orientation of the crystal by measuring the transverse velocities in three or more directions. The

TABLE I. Orientation and path lengths.

	001	110	*hkl*	001	110	*hkl**
φ	~	44½°	126½°	135°	45°	124°
θ	0°	89½°	97½°	1°	90°	83½°
Length	1.092″	0.755″	.936″	.943″	.630″	.967″
	Crystal F-16-a-2			Crystal F-16-a-7		

* The *hkl* surface is near a plane of the form 531 which in turn is near the center of the unit triangle 100-110-111.

TABLE II. Velocity measurements $\times 10^{-4}$ cm/sec.

Mode \ Face	001	110	*hkl*	001	110	*hkl*
V_L'	.3294	.3125	.3135	.3276	.3116	.3119
V_t'	.1056	.1465	.1423	.1055	.1455	.1406
$V_L'V_T'$	.1057	.1056	.1083	.1055	.1054	.1080
	Crystal F-16-a-2			Crystal F-16-a-7		

TABLE III. Elastic constants of silver chloride ($\times 10^{-12}$ dynes/cm^2).

Const. \ Face	001	110	*hkl*	001	110	*hkl*	Ave.
C_{11}	.606	.603	.603	.600	.599	.593	.601
C_{12}	-	.364	.365	—	.362	.356	.362
C_{44}	.0624	.0623	.0627	.0622	.0621	.0631	.0625
$(C_{12}'+C_{44}'+C_{44}')$	.731	.728	.728	.724	.723	.720	.726
$\frac{(C_{11}+2C_{12})}{3}$*	-	.444	.444	—	.441	.436	.441
Dimensionless isotropy ratio	—	1.92	1.90	--	1.90	1.88	1.90
	Crystal F-16-a-2			Crystal F-16-a-7			

* Previously reported values for the bulk modulus are 0.417 reported by Richard and Jones, J. Am. Chem. Soc. 31, 158 (1909) and a value of 0.412 obtainable by extrapolating Bridgman's data to zero pressure.

ratio of the difference of the pairs of transverse waves in directions (a) and (b) is a function of the angles φ and θ alone as:

$$\frac{(V_t'^2-V_T'^2)_a}{(V_t'^2-V_T'^2)_b}=\frac{((B^2+C^2)^{\frac{1}{2}}\sin^2\theta)_a}{((B^2+C^2)^{\frac{1}{2}}\sin^2\theta)_b}. \quad (15)$$

Contours of the $F(\varphi,\theta)=\log(\sin^2\theta(B^2+C^2)^{\frac{1}{2}})$ in stereographic projection have been made to facilitate such operations.

[1] D. L. Arenberg, "Supersonic Solid Delay Lines," M.I.T. Rad. Lab. Report No. 932, February 1946.
[2] J. R. Haynes, Rev. Sci. Inst. 19, 51 (1948). (Note: Use 10 times the total amount of water.)
[3] P. W. Bridgman, Proc. Am. Acad. Arts Sci. 74, 21-25
[4] W. G. Cady, *Piezoelectricity* (McGraw-Hill Book Company, Inc., New York, 1948), p. 83.

Examination of Thin Overgrowths by Multiple Scattering of Electrons

L. G. SCHULZ
Institute for the Study of Metals, University of Chicago, Chicago, Illinois

June 14, 1950

ELECTRON diffraction by the reflection method is commonly employed in the study of oriented overgrowth on single crystals.[1] In Fig. 1A is shown the usual experimental arrangement. When the overgrowth is produced by evaporation and deposition in a vacuum the deposit forming the overgrowth is occasionally too small to be detected easily. To enhance the scattering of very small deposits the angle θ might be decreased. This leads to difficulties, however, because of refraction effects and increased background scattering. It has been found that a better procedure is to employ the reflected beam from the substrate as the incident beam on the deposit. In drawing B of Fig. 1 the part of the incident beam I which is reflected is indicated by I'. Since I' is an incident beam in respect to the deposit there will be a second diffraction process resulting in the beams I''. The angle θ' is associated with the crystal constants of the deposit.

With the arrangement of drawing B one can take advantage of much smaller scattering angles for the deposit than by the direct approach of drawing A. The part of the beam I (in drawing B) which is intercepted by the deposit must be scattered through a total angle of $2\theta\pm\theta'$ to reach the same general area of the photograph as the beams I''. Because of the rapid decrease in intensity with angle of scattering only the spots due to I'' (and to I''') appear in the photograph.

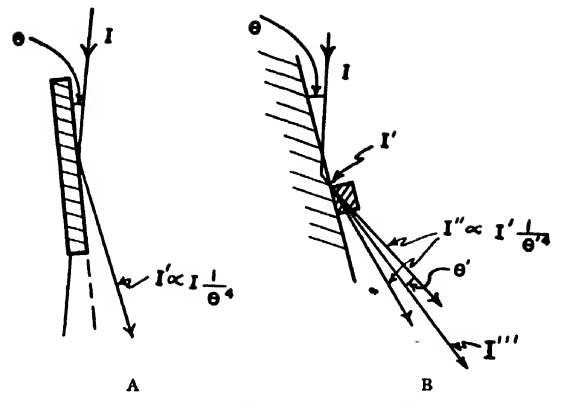
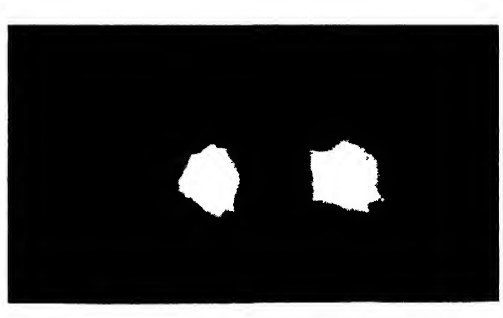


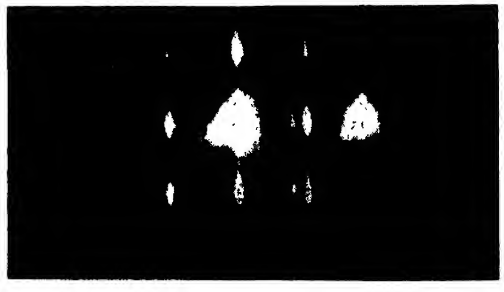
FIG. 1. Experimental arrangements. Drawing A shows the usual reflection arrangement; drawing B shows the new method employing multiple scattering of electrons.

The results from an application to a study in polymorphism are presented in Fig. 2. Photograph (a) is from a single crystal of LiF, the two points being the (600) and (620) reflections. Photograph (b) is from the same crystal but after a deposit of CsBr averaging 2A in thickness has been added. This small deposit has migrated over the substrate and grown into crystals approximately 100A on an edge. An analysis of photograph (b) reveals that the CsBr crystals have a NaCl type structure, rather than the usual CsCl type, and that they are oriented with their axes parallel to the corresponding axes of the substrate.

In Fig. 2 each of the two beams from the LiF substrate is an incident beam in respect to the deposit. As a result there are two secondary diffraction patterns. The displacement of the two patterns relative to each other is the same as the separation of the two LiF points. Consequently the separation of any pair of corresponding points from the two secondary patterns can be used as



(a)



(b)

FIG. 2. Results of an application of the new method. CsBr has been deposited by evaporation and deposition in a vacuum onto a single crystal of LiF. (a): LiF (heavy exposure); (b): 2A of CsBr added to LiF.

an internal standard making possible an accurate determination of the crystal constants of the deposit.

Details concerning the range of application, together with additional results, will be given later.

This research was supported in part by Army Air Force Contract No. AF 33 (638)-6534.

[1] G. P. Thomson and W. Cochrane, *Theory and Practice of Electron Diffraction* (Macmillan and Company, London, 1939), Chapter XI.

A Note on Thermoelectric Generation of Current

MAX B. GREEFF
Upper Montclair, New Jersey
July 5, 1950

TELKES wrote a very thorough and interesting article[1] about the efficiency of a thermoelectric generator. A formula is given for the calculation of the efficiency for any thermocouple whenever the Wiedemann-Franz-Lorenz law is applicable. The calculations given in Table I of the article show that the efficiency of a thermocouple may become much better than has been anticipated in the past, provided we have a couple delivering power beyond 200 microvolts per °C. In fact, at 500 microvolts per °C we may expect an efficiency of 20.9 percent if the temperature gradient between hot and cold junctions is 500°C.

However, there is a limitation in Telkes' formula to which Papet[2] calls our attention. The electrical efficiency, depending on the relation of internal and external resistances, is fixed at 50 percent (internal and external resistances are assumed equal); this being the condition of maximum energy output of the thermoelectric generator. Though we are much concerned in obtaining high energy, it is not always most efficient to operate the generator at maximum energy output.

For example, let us calculate a complete characteristic of a generator assuming 4 different cases of construction: (1) applying thermocouple elements of 100 microvolts per °C; (2) applying thermocouple elements of 500 microvolts per °C; (3) changing the relation of internal and external resistances of case (2); (4) reducing the number of elements in case (3).

The energy output of all 4 cases remains the same (Table I).

TABLE I.

Generator case	*I*	*II*	*III*	*IV*
Power of each couple microvolts per °C	100	500	500	500
Generator e.m.f. volts	100	100	100	55.5
Number of couples required	2000	400	400	222
Voltage utilized volts	50	50	95	50
Amperage utilized amp.	25	25	13.2	25
Wattage utilized kilowatts	1.25	1.25	1.25	1.25
Resistance intern. ohms	2	2	0.4	0.22
Resistance extern. ohms	2	2	7.2	2.00
Resistance total ohms	4	4	7.6	2.22
Resistance per couple ohms	0.001	0.005	0.001	0.001
A efficiency Telkes %	2.34	20.9	—	—
B electrical $\frac{r_2}{r_1+r_2}$ %	50	50	95	90
C eff. total $=2A \times B$ %	2.34	20.9	39.8	37.6

We see immediately that there is more than one way to construct a thermoelectric generator of a given output. At maximum output, Cases *I* and *II*, we may adopt the calculated efficiencies offered by Telkes. But Cases *III* and *IV* require discussion as they are not at maximum output. Consequently, we are not allowed to apply Telkes' efficiency for these cases. We may, however, compare the two cases *III* and *IV* with Case *I* which has the same resistance per element and only differs in the number of elements.

Case *III* has 1/5 the number of elements as has Case *I*, producing the same energy. Consequently, Case *III* has 5 times the efficiency as Case *I*, using only 1/5 the amount of fuel. The efficiency therefore is $5 \times 2.34 = 11.60$ percent.

Case *IV* has 1/9 the number of elements as has Case *I*, producing the same energy. Consequently, Case *IV* has 9 times the efficiency as Case *I*, using only 1/9 the fuel. Therefore, the efficiency is $9 \times 2.34 = 21.06$ percent.

We could construct a generator of the same elements and equal output, but at different voltage, whose efficiency is still higher, choosing the special Case *V* where the numerical figures for voltage and amperage are practically equal.

This Case *V* would require only 160 elements or thermocouples to deliver a voltage of 36 volts at an amperage of 35 amp. Using only 1/12.5 of the number of elements as in Case *I*, the efficiency is $12.5 \times 2.34 = 29$ percent. Under these circumstances the generated e.m.f. would be 40 volts and the electrical efficiency 90 percent.

[1] M. Telkes, J. App. Phys. 18, 1116 (1947).
[2] R. M. Papet, J. App. Phys. 19, 1180 (1948).

On the Negligible Changes Produced in the Stress-Strain Curve by Immersion of a Copper Single Crystal in Mercury

J. B. WACHTMAN, JR.
Carnegie Institute of Technology, Pittsburgh, Pennsylvania
July 6, 1950

IN his proposed explanation of slip bands, Frank[1] considers internal processes and reflections of fast dislocations at the surface of a single crystal as dislocations of opposite sign. He suggests that immersing a single crystal in a liquid of higher density will suppress reflections of dislocations. If reflection of dislocations plays an appreciable role as compared with internal processes in the formation of slip bands, one might expect the suppression of reflections to change the stress-strain curve.

Single crystals of 99.999 percent pure copper were grown to A.S.T.M. standard tensile specifications (2-in. gauge length, 0.25-in. diameter). The crystals were oriented and then annealed in vacuum for 20 hours at 600°C.

These samples were then subjected to tensile tests in a Baldwin Southwark Tate-Emery 60,000-pound testing machine; the extension being determined from the travel of the head of the testing machine as determined by a dial gauge. This method of extension measurement was checked on three crystals by measuring the separation of two marks on the crystal initially 1 inch apart with two cathetometers. The stress and strain were resolved by assuming that single slip took place until a position of symmetry with respect to two slip systems was reached after which it was assumed that equal slip on both systems took place. Three crystals were pulled partly in air and partly under mercury. Table I gives, for each crystal, the resolved shear strain at which double slip begins, the resolved shear strain at which mercury was added, and the final load.

TABLE I.

Crystal No.	Final load	Resolved shear strain at beginning of double slip	Resolved shear strain at which mercury was added
3	1022 lb.	53.8%	No Hg added
4	764	48.4%	9.5%
6	768	29.5%	No Hg added
10	1012	52.1%	No Hg added
11	731	56.4%	0.0%
13	823	64.0%	10.3%

The resolved stress-strain curves for the crystals strained in air are shown in Fig. 1 as calculated from the cathetometers and the dial gauge extensometer. The two methods of calculation yield curves in good agreement with each other for crystals 3 and 10. The curve for No. 6 computed from the dial gauge has an unusual bump in it and does not agree well with the other curve for No. 6 at intermediate strains. The strain determined from the

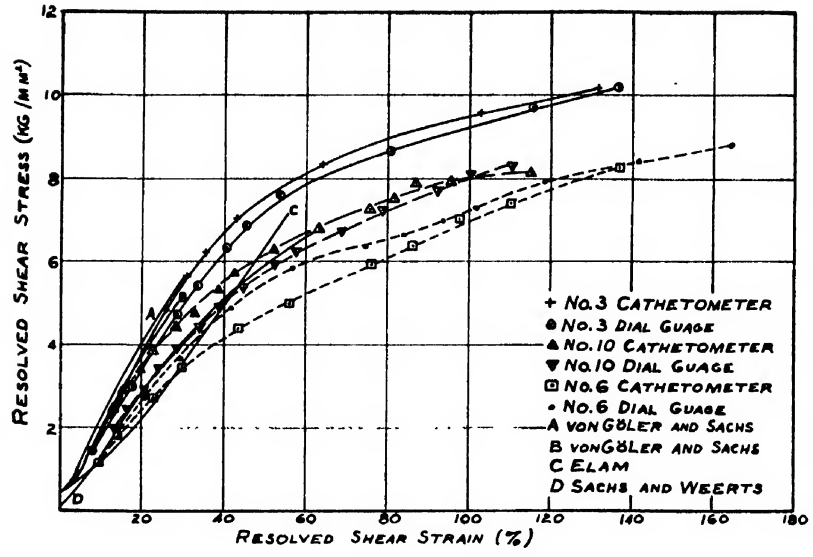


FIG. 1. Stress-strain curves obtained on copper crystals in air.

dial gauge may be in error because of inhomogeneous deformation near the grips, slipping of the grips in the machine, or sticking of the plunger in the dial gauge. Since the high strain data yields close agreement it seems reasonable to assume that the error at intermediate strain was caused by sticking of the dial gauge plunger and that the high and low strain portions of No. 6 curve, in agreement with No. 3 and No. 10, indicate that inhomogeneous deformation near the grips and slipping of the grips do not produce large errors in the strain computed from the dial gauge.

The curves for No. 4 and No. 13 lie well within the range of the curves run in air. The curve for No. 11 does drop below the lowest of the curves run in air but the greatest stress difference between it and the curve for No. 6 is 0.8 kg/mm² which is to be compared with a spread of 2.9 kg/mm² between the curves run in air. Thus if the weakness of No. 11 is ascribed to the mercury (which is improbable in view of the curves for No. 4 and No. 13) the effect is smaller than the normal variation in annealed, high purity crystals.

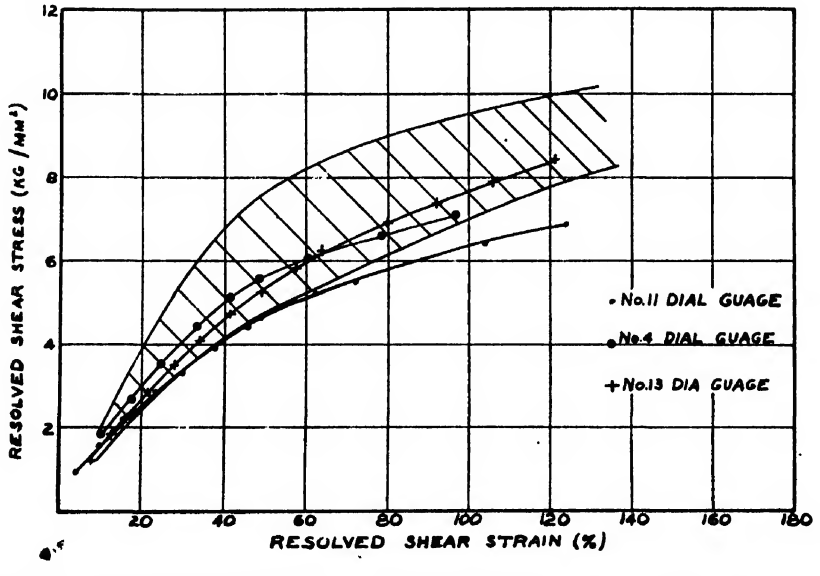


FIG. 2. Comparison of stress-strain curves obtained in air and in mercury. The shaded region gives the spread of the curves obtained in air.

Curves by von Goler and Sachs,[2] by Sachs and Weerts,[3] and by Elam[4] are also shown in Fig. 1.

The curves obtained from the mercury experiments are shown in Fig. 2. The shaded region indicates the spread of the curves obtained on specimens run in air.

The curvature and the roughness of the surface of the crystals made a microscopic examination of the slip lines impossible but to the naked eye the appearance of the slip lines was the same on crystals strained under mercury and those strained in air.

In conclusion it appears that immersion of copper single crystals in mercury does not have any important influence on the resolved stress-strain curve.

The author wishes to express his gratitude to Dr. J. S. Koehler for his aid.

[1] F. C. Frank, *Report of a Conference on Strength of Solids.*
[2] F. von Goler and G. Sachs, Zeits. f. Physik **55**, 581 (1929).
[3] G. Sachs and J. Weerts, Zeits. f. Physik **62**, 473 (1930).
[4] C. F. Elam, Proc. Roy. Soc. **A112**, 289 (1926).

Journal
of
Applied Physics

Volume 21, Number 10 October, 1950

The Effect of Input Configuration on Antenna Impedance*

J. R. Whinnery
University of California, Berkeley, California
(Received February 13, 1950)

It is known that the configuration of the input driving system may be of importance to the impedance of antennas at high frequencies. In this paper, a region in the vicinity of the input is considered as a transducer between the TEM mode on the driving transmission line and the spherical nearly-TEM mode on the antenna system, and the constants of this transducer are found both by calculation and measurement for a particular case. The calculation is based upon quasi-static concepts, utilizing an electrostatic field map obtained from an electrolytic tank to give the approximate form of the electric field lines. The measurement is made by finding three input reactances corresponding to the input region closed by three spun aluminum hemispheres of different radii. The three sets of data, taken at each frequency, are sufficient to determine the constants of the transducer. They are expressed here as a pi-network for comparison with calculated values.

The main portion of the antenna is analyzed substantially as in Schelkunoff's approach, with the TM modes in space and in the antenna region expressed as a shunt admittance at the end of the antenna in the TEM mode equivalent circuit. This admittance is then transformed to the desired reference on the antenna by a perturbation calculation. From there it is transformed to a desired reference on the feed line by means of the input network constants. Values calculated in this way are compared with measured values over the range 200 to 1000 mc/sec. for the particular monopoles studied. Agreement is not perfect, but appears considerably better than without consideration of the specific input configuration.

INTRODUCTION

IT is true that in many practical antenna problems, especially at frequencies below 100 mc/sec., the actual input driving means may add only a small contribution to input admittance. In these cases, it is not necessary to consider the actual input configuration in a mathematical analysis of the antenna, but, if an infinitesimal driving gap is assumed, the infinite susceptance singularity which would ordinarily result from this assumption must be avoided in some way. It will be described in the following paragraphs how this is accomplished in three of the modern methods of antenna analysis.

For frequencies above 100 mc/sec. (and for some antennas below this range), the actual input configuration may be important to the input impedance of an antenna. If the local capacitance of the actual input system is known, a first approximation may be had by adding the corresponding susceptance to the input admittance calculated by any of the analyses which have removed the local effects by avoiding the singularity. As an example, in the measured curves of Brown and Woodward,[1] their Fig. 13 shows curves of input resistance *versus* antenna length for antennas identical except with respect to their input configurations. These curves differ widely. They were able to explain much of the difference in these results in terms of an effective shunt capacitance across the input. However, it is clear that for many practical antenna inputs (for example some of those sketched in Fig. 1 for driving a monopole antenna above plane ground from a coaxial line), the effect of the input is too complex to be represented by a simple shunt capacitance, and a more general study of the input configuration is desirable.

It is the purpose of this paper to describe a point of view that may be applied to the input configuration of certain antennas, with examples of its use. The approach, especially that of the quasi-static method to be described, is applicable to a variety of boundary-value problems, such as those of discontinuities in wave guides.

* This work was done at the Antenna Laboratory of the University of California under Bureau of Ships Contract NObsr-39401, and constitutes a portion of a thesis submitted in partial fulfillment for the degree of Ph.D. at the University of California.

[1] G. H. Brown and O. M. Woodward, Proc. I.R.E. 33, 257–262 (April, 1945).

THE SOURCE PROBLEM IN STANDARD ANTENNA ANALYSES

In the analysis of forced oscillations on spherical[2] and spheroidal[3] antennas, Stratton and Chu summed series of the appropriate wave solutions to fit the boundary conditions of the conducting antenna with an infinitesimal gap at its center across which the driving voltage is applied. The expected infinite capacitance results, since the series for susceptance are observed to diverge. The authors, in commenting on this, have pointed out that, for a finite gap, the amplitudes of

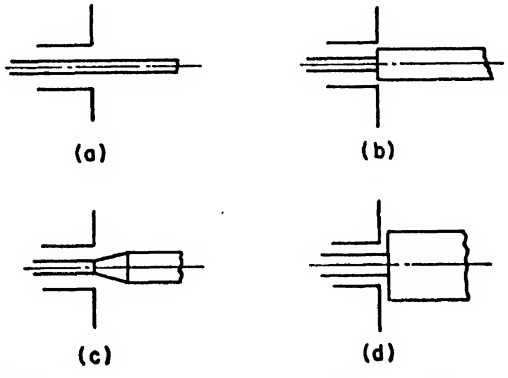


FIG. 1. Various input configurations for monopole antennas.

wave orders in the series would eventually approach zero as an order is reached great enough to give many reversals in sign in the angular function over the extent of the gap. They therefore recommend terminating the susceptance series after a finite number of terms to correspond to some actual gap. Infield[4] has made some estimates of the number of terms to retain for a given allowed error.

In Schelkunoff's analysis of the ideal biconical antenna,[5] the gap problem causes no mathematical difficulty, for, since the source applied between the tips of the two cones is a point source and not a "ring" or "slice" source as in the corresponding cylindrical dipole, there is no singularity in the admittance function. In his extension of results to thin antennas of other shapes, for example the cylindrical dipole of Fig. 2, it would appear that the original gap problem would return. Schelkunoff has a carefully outlined procedure for considering the specific gap,[6] but, even in the direct application of his perturbation solution with antenna admittance computed exactly at the origin, the infinite susceptance does not appear, and the calculated results agree well with measured values over the range of antennas for which input configuration is not important.

In order to explain the avoidance of the singularity in the Schelkunoff perturbation calculation, consider its application to the cylindrical antenna of Fig. 2. Each element of the antenna is considered first as an element of the biconical antenna passing through that radius, and the characteristic impedance computed for that point from the biconical line formula.

$$K = 120 \ln \cot(\psi/2) \cong 120 \ln[2z/\rho(z)]. \qquad (1)$$

There are at least two steps in this approximation, the neglect of the finite slope of the cone passing through $\rho(z)$ in comparison with the zero slope of the cylinder, and the substitution of $2z/\rho(z)$ for $\cot(\psi/2)$. Both of these are valid for large values of $z/\rho(z)$, and these approximations enter into the limitations of the method to thin antennas. However, the approximations also fail near the origin for antennas of any size, and they fail in just such a way that the infinity of susceptance is avoided. Thus the local capacitance effects and the susceptance effects from the antenna proper are separated by this process, and the resulting admittance is satisfactory for the large number of practical cases for which the local input effects are of small importance to the antenna impedance.

For consideration of the effects of a finite gap, Schelkunoff[6] recommends that one transfer the terminating admittance obtained from his biconical line analysis only to such a radius from the origin that the approximations described above (requiring that $z/\rho(z) \gg 1$) are valid, but also small enough compared with wave-length so that quasi-static ideas may be used in this region. This admittance is then transformed to $z=0$ by uniform line equations corresponding to the biconical antenna passing through this radius. To the resulting finite admittance, one adds the difference between the capacitive susceptance of the actual cylinder, including its physically real gap, and that of the imaginary biconical antenna utilized in the above step to avoid the singularity at the origin. The method to

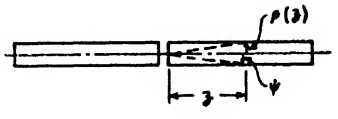


FIG. 2. Basis for Schelkunoff's formulation for other than biconical antennas.

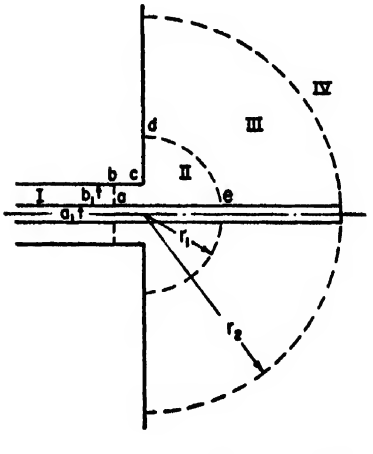


FIG. 3. Division of the antenna problem.

[2] J. A. Stratton and L. J. Chu, J. App. Phys. 12, 236–240 (1941).
[3] L. J. Chu and J. A. Stratton, J. App. Phys. 12, 241–248 (1941).
[4] L. Infield, Quart. App. Math. 5, 113–132 (July, 1947).
[5] S. A. Schelkunoff, Proc. I.R.E. 29, 493–521 (Sept. 1941).
[6] S. A. Schelkunoff, J. App. Phys. 15, 54–60 (1944).

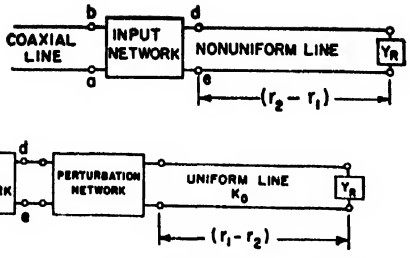


FIG. 4. Equivalent circuit for antenna of Fig. 3.

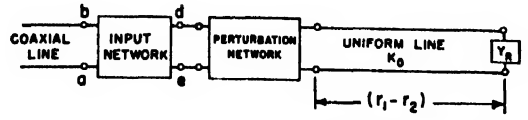


FIG. 5. Alternative equivalent circuit for the antenna of Fig. 3.

be described in this paper is in many respects an extension of this recommended procedure to other input configurations.

A third important method for the analysis of antenna impedance is that based upon approximate solutions of the integral equation developed by Hallen.[7-9] Since this analysis assumes the source as an infinitesimal gap in the cylinder across which there is a discontinuity in the scalar potential equal to the applied voltage, it would be expected to encounter the infinite susceptance difficulty described. However, it is common in these analyses to make an approximation to the angular integration in computing the vector potential, which amounts to assuming the current concentrated along the axis for purposes of computing the potential. This in effect evidently replaces the "slice" source at the gap by a point source, and so avoids the singularity as in the previous analyses.

THE POINT OF VIEW OF THIS PAPER

The point of view to be followed in this paper is: first, that the antenna system should be considered as a transducer from the feeding line to the waves in space as in Schelkunoff's analysis; secondly, that the problem should be divided into two parts, one consisting of the main body of the antenna to be analyzed substantially as in Schelkunoff's approach, and the other consisting of a region near the input, taking account of the actual input configuration. Considering specifically the monopole above ground plane, fed by a coaxial line (Fig. 3), one imaginary spherical surface is drawn through the end of the antenna at radius r_2 separating regions III and IV, and the matching of spherical wave functions across this surface is accomplished as in Schelkunoff's analysis, or by other methods. The effect of the TM modes in regions III and IV is expressed as a shunt admittance at the end of the antenna in the principal wave equivalent circuit, and this admittance is transformed to radius r_1, say by the non-uniform line formulation of Schelkunoff. All this is as in Schelkunoff's recommended procedure described above. However, at this stage, the entire region II between a reference plane $a-b$ in the coaxial line and the imaginary spherical surface $d-e$ of radius r_1 is considered as a two-terminal-pair network or transducer** whose parameters may be found either by calculation or by measurement. Once the parameters are known, and expressed in any of the possible forms (T networks, pi networks $\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}$ constants, etc.), the admittance may be simply transformed from $d-e$ at radius r_1 to the reference $a-b$ in the coaxial line, yielding the desired information.

The reference plane $a-b$ is first selected far enough from the end of the line so that the local waves from the discontinuity have died out and the electric field is substantially radial; similarly the radius r_1 is selected far enough from the discontinuity so that the electric field lies substantially in the spherical surface $d-e$. This is the basic assumption of the approach. There could fail to be any such spherical surface if the antenna were too fat, or if it were not long enough compared to the dimensions of the input. That is, it is desired to select r_1 so that it satisfies the inequalities

$$r_1 \gg b_1 \tag{2}$$

$$r_1 \ll r_2, \tag{3}$$

where the $\gg$ is interpreted in this case only as "appreciably greater than." A reasonable selection of r_1 could thus be found easily for most antennas up to 1000 mc/sec., but not for most microwave antennas. In addition to the above inequalities, it is usually possible to choose r_1 to satisfy

$$r_1 \ll \lambda, \tag{4}$$

which is not essential, but does simplify the analysis.

The assumption that there is some region over which the electric field lines are substantially spherical, so that the analyses of the input configuration and the discontinuity at the end of the antenna may be separated, is much like that encountered in wave-guide discontinuity problems. If there are two discontinuities in a

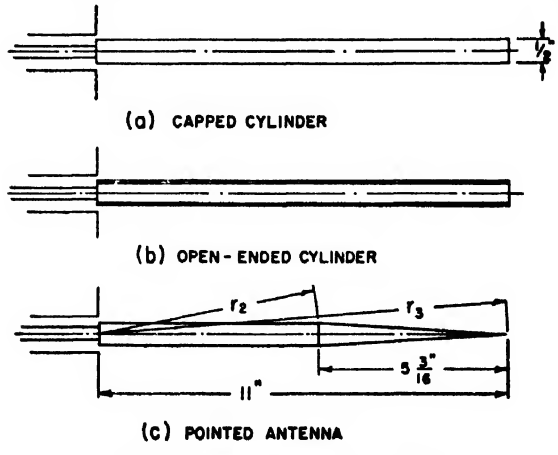


FIG. 6. The three antennas studied.

[7] E. Hallen, Nova Acta (Uppsala) 11, 1-44 (November, 1938).
[8] C. J. Bouwkamp, Physica 9, 609-631 (July, 1942).
[9] R. King and D. Middleton, Quart. App. Math. 3, 302-335 (January, 1946).

** For the application of classical network theory to general electromagnetic structures, see for example, Montgomery, Dicke, and Purcell, *Principles of Microwave Circuits*, M.I.T. Rad. Lab., Vol. 8 (McGraw-Hill Book Company, Inc., New York, 1948).

wave guide far enough apart so that the local waves from each have effectively died out in some intermediate region, the dominant mode only existing in this region, the analysis of the two may proceed separately, the network for each found when alone in the guide, and the results of both found by combining by means of the transmission line equations. If they are so close that there is no such region, then they cannot be so separated.

The over-all equivalent circuit for the antenna (Fig. 4) will then consist of the input network representing region II, the non-uniform line representing the region III, and the terminating admittance Y_R representing the radiation and discontinuity effects at the end of the antenna. As will be shown later, it is possible to represent the non-uniform line by a uniform line of average characteristic impedance with a perturbation network at its input, so that the over-all equivalent circuit would then become as in Fig. 5. Of course, although the reference plane $a-b$ in the coaxial line should first be selected in the analysis far enough from the discontinuity so that there is only the principal TEM mode, and all measurements should be made at least this far from the end effect, the reference on the transmission line equivalent circuit can then be shifted at will by ordinary transmission line transformations. This will almost always be done, shifting the reference to the end of the line, since it forms the most natural reference.

THE SPECIFIC PROBLEMS STUDIED AS EXAMPLES

The point of view outlined above could be applied to a number of input configurations, such as those of Fig. 1. The particular configuration chosen for study was that of Fig. 1b, in which a $\frac{3}{4}$-in. O.D., $\frac{1}{4}$-in. I.D. coaxial line feeds a $\frac{1}{2}$-in. diameter cylindrical antenna with an abrupt step between the center conductor of the coaxial line and the antenna. This input structure was used on the three different antennas pictured in Fig. 6, each eleven inches long and $\frac{1}{2}$ inch in diameter at the base. That of Fig. 6(a) is a circular cylinder closed by a plane cap at the end, that of Fig. 6(b) is a cylinder left open at the end, and that of Fig. 6(c) is tapered in the last $5\frac{3}{16}$ inches, forming a cone with its apex at the antenna end. The input impedance for each was calculated, including the effects of the input configuration by means of the equivalent input network, and compared with measured values for a frequency range of 200 to 1000 mc/sec.

A SECOND-ORDER QUASI-STATIC CALCULATION OF THE INPUT NETWORK

In order to fulfill simultaneously the two inequalities of Eqs. (2) and (3) for the $\frac{1}{2}$-inch diameter antenna,

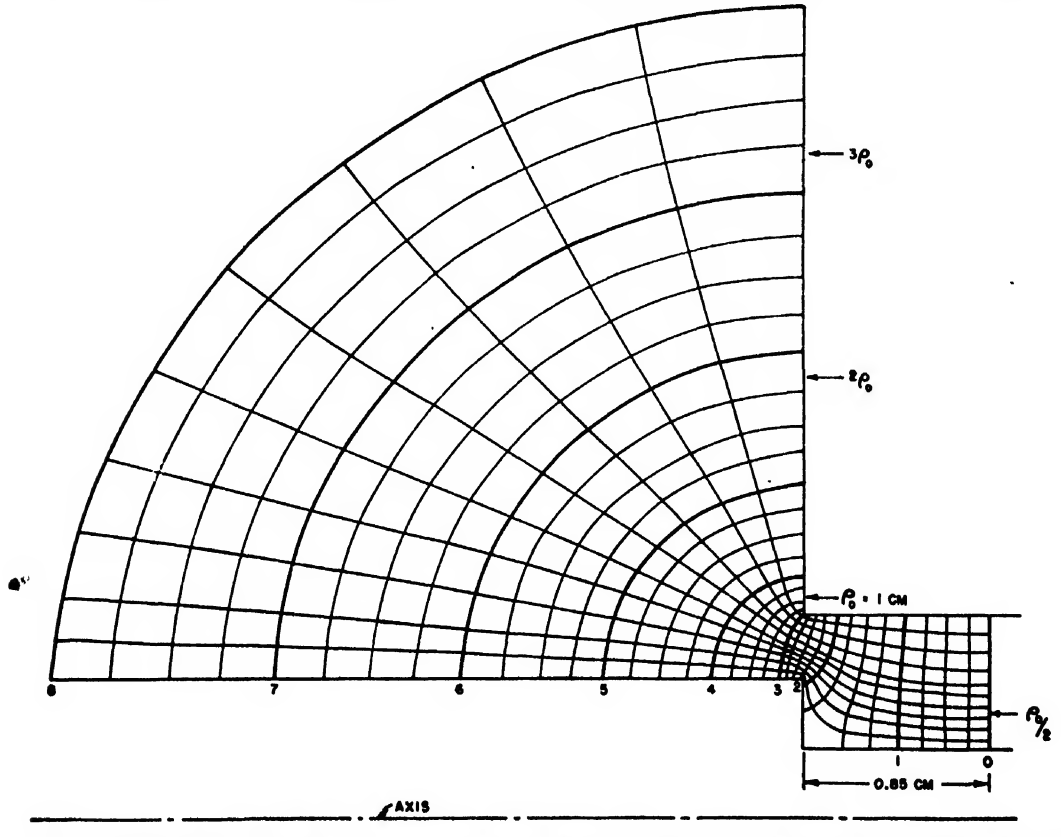


FIG. 7. Electrostatic field map of the antenna input region. Heavy flux lines indicate the divisions for the application of the second-order quasi-static analysis.

eleven inches long, fed by the $\frac{3}{4}$-inch (outer conductor inside diameter) coaxial line, the output reference sphere at radius r_1 was chosen as 3.45 cm. This yields $r_1/b=3.63$ and $r_2/r_1=8.1$, which is believed to be a satisfactory fulfillment of these inequalities. The inequality of Eq. (4) is also satisfied over the range of interest since the shortest wave-length (corresponding to 1000 mc/sec.) is 30 cm. An analysis by quasi-static methods is then appropriate for approximate determination of the input network.

A first-order equivalent circuit may be obtained from quasi-static ideas by determining for the region a series inductance L and a shunt capacitance C. The way in which the capacitance is divided between the branches of a pi (or the inductance between the arms of a T) affects the transforming properties of the network only to terms of the order ω^2LC. As is shown in Appendix A, the quasi-static ideas may be extended to give the proper division of the capacitance between the branches of the pi network for low frequencies. This is useful because the elements of the pi network actually remain substantially constant over a wide range of frequencies. For example, the experimental results show that the inductance and capacitances of the pi network representing the input configuration of the present problem change little up to 1500 mc/sec., which is considerably beyond the range of interest in this study.

The method is a quasi-static method in that it utilizes the electrostatic field map for the region, and so assumes that it is proper to approximate the wave equation by Laplace's equation. The electric field map utilized was obtained from an electrolytic tank, tilted about an axis as required for axially symmetric problems, utilizing tap water as an electrolyte, and excited by a 1000-c.p.s. signal generator. The field map obtained is shown in Fig. 7. The over-all region is subdivided by the flux tube boundaries, as indicated in Fig. 8, into regions small enough so that each of these may be represented by a symmetrical T. These are then cascaded to represent the over-all region. In the calculation of the matrix of the over-all network, terms of the order ω^2LC are retained, but higher powers of $(\omega[LC]^{\frac{1}{2}})$ are neglected. As is shown in Appendix A, results can be expressed by the pi-network of Fig. 9 with

$$L=\sum_{k=1}^{N} L_k, \tag{5}$$

$$C_a=C_n\sum_{k=1}^{N}(k-\tfrac{1}{2})\frac{L_k}{L}, \tag{6}$$

$$C_b=C_n\sum_{k=1}^{N}(N-k+\tfrac{1}{2})\frac{L_k}{L}. \tag{7}$$

The capacitance of the nth tube may be obtained from the electrostatic field map in the usual manner.

$$C=2\pi\epsilon\rho_0(N_f/N_v), \tag{8}$$

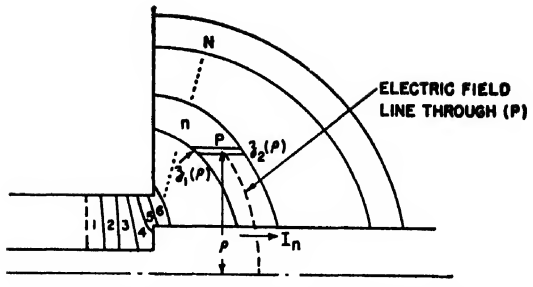
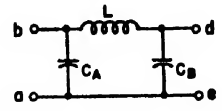


FIG. 8. Construction indicating the nth flux tube of the electrostatic field map of the region.

FIG. 9. Pi-network representing the over-all input region.



where ρ_0=radius at which side ratios of curvilinear rectangles is made unity; N_f=number of flux tubes; N_v=number of potential divisions; ϵ=dielectric constant$\cong$8.85 $\mu\mu$f/meter for air or space. The inductance of the nth tube may be obtained by graphically integrating the magnetic flux over the section of the nth tube. The magnetic field for this purpose may be related to the current in the nth tube. Thus if the line integral of magnetic field is evaluated at radius ρ passing through any point P in the nth tube (Fig. 8), it can be set equal to the current I_n, since a surface through P following an electric field line (shown dashed) would have no displacement current passing through it. So

$$H_\phi=I_n/2\pi\rho.$$

Here the current in the individual subdivision is assumed constant, and this can be satisfied by making the divisions small enough. The inductance of the nth tube is then

$$L_n=\frac{1}{I_n}\int\int \mu H_\phi d\rho dz=\frac{\mu}{2\pi}\int\int_{\text{tube section}}\frac{d\rho dz}{\rho}. \tag{9}$$

Either the ρ or z integration may be performed so that only one needs to be made graphically. Performing the z integration,

$$L_n=\frac{\mu}{2\pi}\int_{\rho\,\min}^{\rho\,\max}\frac{[z_2(\rho)-z_1(\rho)]}{\rho}d\rho, \tag{10}$$

where $z_2(\rho)$ and $z_1(\rho)$ represent the limits in z represented by the tube at a given value of radius.

For application of the method to the present problem, the region represented by the electrostatic map of Fig. 7 was divided into eight sub-regions, indicated by the heavy lines. With $N_v=10$, $N_f=4$, and $\rho_0=1$ cm, the capacitance per division from Eq. (8) was calculated as 0.222 $\mu\mu$f. The graphical integration for computation of the inductance according to Eq. (10) yielded

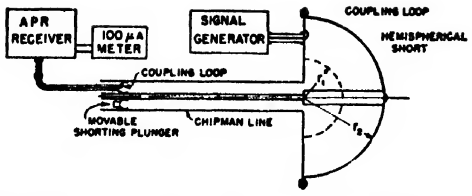


FIG. 10. Test set-up for measurement of the input region.

the following results:

$$L_8=4.14\times10^{-9}\text{ henrys}$$
$$L_7=3.03\times10^{-9}\text{ henrys}$$
$$L_6=1.88\times10^{-9}\text{ henrys}$$
$$L_5=0.954\times10^{-9}\text{ henrys}$$
$$L_4=0.318\times10^{-9}\text{ henrys}$$
$$L_3=0.153\times10^{-9}\text{ henrys}$$
$$L_2=0.858\times10^{-9}\text{ henrys}$$
$$L_1=0.865\times10^{-9}\text{ henrys.}$$

Thus, using Eqs. (5)–(7)

$$L'=\sum_{k=1}^{8}L_k=12.19\times10^{-9}H, \tag{11}$$

$$C_A'=\frac{0.222}{12.19}\left[\frac{L_1}{2}+\frac{3L_2}{2}+\cdots\frac{15L_8}{2}\right]=1.250\ \mu\mu\text{f}, \tag{12}$$

$$C_B'=\frac{0.222}{12.19}\left[\frac{15L_1}{2}+\frac{13L_2}{2}+\cdots\frac{L_8}{2}\right]=0.527\ \mu\mu\text{f}. \tag{13}$$

The values of Eqs. (11)–(13) are primed because it is now desired for convenience to move the reference in the transmission line to the end of the line by subtracting out the part of the network arising from the transmission line's *principal* wave. Note that this still leaves in the network the effects of the higher order waves in the transmission line region. Writing the above network as the cascade connection of two networks, the desired one and the network representing the principal transmission line wave over the length 0.85 cm, which by usual formulas yields values of $L_0=1.87\times10^{-9}$ henry and $C_0=0.436\times10^{-12}$ Fd, the approximate elements of the desired network can be found as follows:

$$L=L'-L_0=10.32\times10^{-9}\text{ henry}, \tag{14}$$

$$C_a=\frac{L'}{L}C_A'-C_0\left(1+\frac{L_0}{2L}\right)=1.005\ \mu\mu\text{f}, \tag{15}$$

$$C_b=C_B'-\frac{L_0}{L'}\left(C_A+\frac{C_0}{2}\right)=0.340\ \mu\mu\text{f}. \tag{16}$$

EXPERIMENTAL DETERMINATION OF THE INPUT NETWORK

One of the common methods for determining the constants of a network for representation of wave-guide discontinuities utilizes the measurement of input admittances Y_{i1}, Y_{i2}, Y_{i3}, corresponding to known load admittances Y_{L1}, Y_{L2}, Y_{L3}. The admittance parameters for the network may then be calculated from the equations

$$Y_{11}=\frac{(Y_{i1}-Y_{i3})(Y_{i1}Y_{L1}-Y_{i2}Y_{L2})-(Y_{i1}-Y_{i2})(Y_{i1}Y_{L1}-Y_{i3}Y_{L3})}{(Y_{i1}-Y_{i3})(Y_{L1}-Y_{L2})-(Y_{i1}-Y_{i2})(Y_{L1}-Y_{L3})} \tag{17}$$

$$Y_{22}=\frac{(Y_{i1}Y_{L1}-Y_{i2}Y_{L2})-Y_{11}(Y_{L1}-Y_{L2})}{Y_{i2}-Y_{i1}} \tag{18}$$

$$Y_{12}^2=(Y_{11}-Y_{ik})(Y_{22}+Y_{Lk}),\quad k=1,2,3. \tag{19}$$

The known load admittances are often obtained by placing shorts on the output guide, and the input admittances measured by determining the standing wave ratio and position of a voltage node on the input guide. If losses in the system are negligible, the admittances are pure susceptances, and all Y's may be replaced by the corresponding B's (where $Y=jB$) in Eqs. (17)–(19).

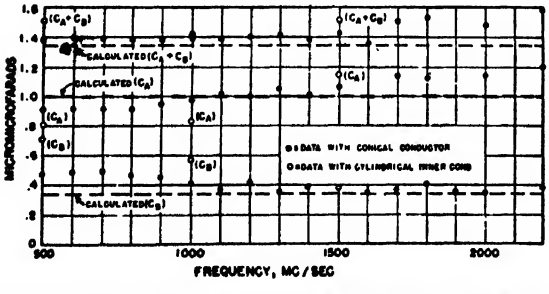


FIG. 11. Comparison of calculated and measured results for capacitances of the input network.

A similar procedure was followed in the experimental determination of the network for representation of the antenna input region in the present investigation. To produce known terminating susceptances at the reference radius r_1 chosen as the output reference of the network, hemispherical shorts to the spherical TEM mode were placed at a radius $r_2>r_1$, as indicated in Fig. 10. The positions of the current maxima on an input coaxial line were observed, giving the corresponding susceptances at the input reference. This procedure was repeated at each frequency with four different hemispherical shorts of nominal diameters 6, 9, 12, and 15 inches, the fourth pair of input-output data providing a check on the three pairs required for the calculation. Most of the measurements were made with

the inner conductor flared conically between r_1 and r_2 in order that a true spherical TEM mode might exist in this region and B_L might be computed in terms of the short position by uniform transmission line equations. It is true that this introduces a discontinuity not present on the antennas themselves, but it is believed that the effect will be small because of the smallness of the flare required. A few measurements were made for a check with the input conductor maintained cylindrical, but in these the data had to be reduced by the longer non-uniform line calculations. The test set-up for all measurements is as indicated in Fig. 10.

The admittance produced at r_1 by the length of shorted spherical TEM mode is

$$Y_L = jB_L = -j\left[\frac{1}{60 \ln \cot(\psi/2)}\right] \cot\frac{2\pi}{\lambda}(r_2 - r_1), \quad (20)$$

ψ = the cone angle, measured from the axis.

Once the admittance parameters are calculated from three pairs of input-output data by Eqs. (17)–(19), the elements of the pi-network may be found from the equations

$$\omega C_a = B_{11} - B_{12}, \quad (21)$$

$$\omega C_b = B_{22} - B_{12}, \quad (22)$$

$$\omega L = -1/B_{12}. \quad (23)$$

It has been mentioned that four hemispheres were constructed to give an additional pair of input-output data for a check, and that some check points were taken with the cylindrical inner conductors. The consistency check for the extra data was made by means of a graphical construction described by Weissfloch[10] (reactive transformation diagram). The three pairs required for the calculation from Eqs. (17)–(19) were selected as those having the greatest difference in admittance values, once unreliable points had been eliminated by the Weissfloch diagram.

The results for C_a and C_b (and for C_a+C_b) are shown in Fig. 11 compared with the dashed line showing the calculated result of Eqs. (15), (16). The comparison for the inductance L is shown in Fig. 12. Although agreement is not perfect, it is believed sufficient to check the principle used. Aside from the fact that the method of calculation was an approximate one, much of the difference is believed to be caused by the fact that the hemispheres could not be spun perfectly, so that they were more nearly spheroids with major and minor half-axes differing by as much as one-fourth inch. Uncertainty in reading the plunger position, although only of the order 0.1 mm, could also explain some of the deviation. A study of the equations revealed that the 0.1 mm deviation could account for as much as a three percent error in the elements of the circuit.

[10] A. Weissfloch, Hoch:tech. u. Elek:akus. 61, 100–123 (April, 1943).

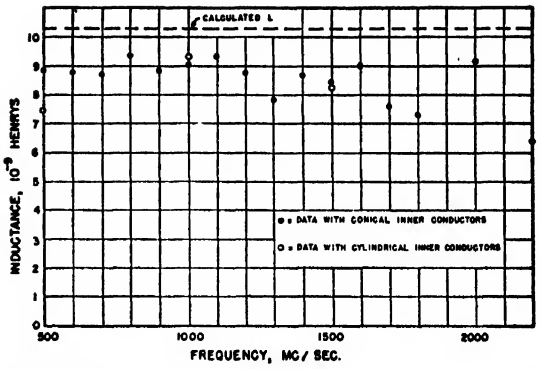


FIG. 12. Comparison of calculated and measured results for inductance of the input network.

NON-UNIFORM LINE REPRESENTATION OF THE ANTENNA

Schelkunoff's analysis of thin antennas of other than conical shape utilizes a perturbation solution which may be interpreted in terms of transformation of impedances along a non-uniform transmission system, as described earlier. As is shown in Appendix B, the Carson-Schelkunoff perturbation solution for slightly non-uniform lines may be described by stating a "perturbation" network to be connected at the input of the corresponding uniform line of average characteristic impedance and propagation constant. The $\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}$ constants of the perturbation network are given in Eqs. (47) and (45) of Appendix B. For the antenna problem, each element of the antenna is represented by a section of the uniform spherical TEM mode, so the special forms of Eqs. (48)–(49) for a loss-free system with constant LC product at each reference on the system are applicable.

$$\mathfrak{A}_1(l) = 1 - \frac{M(l)}{K_0} \quad (24)$$

$$\mathfrak{B}_1(l) = -jN(l) \quad (25)$$

$$\mathfrak{C}_1(l) = j\frac{N(l)}{K_0^2} \quad (26)$$

$$\mathfrak{D}_1(l) = 1 + \frac{M(l)}{K_0}, \quad (27)$$

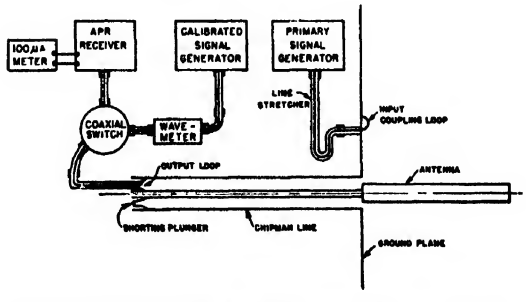


FIG. 13. Set-up for antenna impedance measurements.

where

$$M(l)=\beta\int_0^l [K_0-K(z)]\sin 2\beta z dz \tag{28}$$

$$N(l)=\beta\int_0^l [K_0-K(z)]\cos 2\beta z dz \tag{29}$$

$K(z)$ = the local characteristic impedance for a given reference z

$$K_0=\frac{1}{l}\int_0^l K(z)dz \tag{30}$$

β = phase constant, $\omega(LC)^{\frac{1}{2}}$.

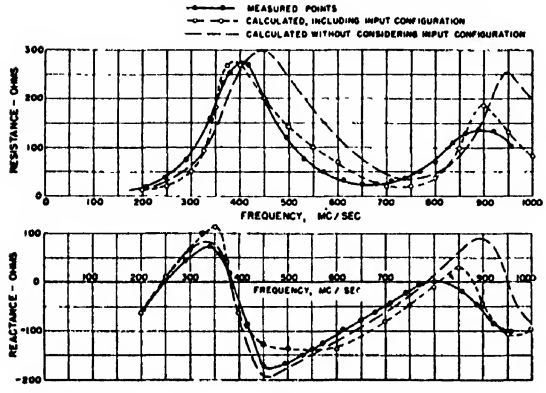


FIG. 14. Input impedance of cylindrical antenna.

For the cylindrical antenna, the equations for M, N, and K_0 given by Schelkunoff[11] must be modified slightly since it is desired to begin the nonuniform line not at the origin, but at a finite radius r_1. Thus

$$z\cong r-r_1.$$

Eq. (1) for the local characteristic impedance $K(z)$, approximated from that for the spherical TEM mode for this radius, becomes

$$K(z)\cong 60\ln\left[\frac{2r}{\rho(r)}\right]=60\ln\left[\frac{2(r_1+z)}{a}\right]. \tag{31}$$

Substitution of this value in Eqs. (28)–(30) yields

$$K_0=60\left[\frac{r_2}{l}\ln\frac{2r_2}{a}-\frac{r_1}{l}\ln\frac{2r_1}{a}-1\right] \tag{32}$$

$$M(l)=30\left\{\ln\frac{r_2}{r_1}+(1-\cos 2\beta l)\left[\frac{r_1}{l}\ln\frac{r_2}{r_1}-1\right]-\cos 2\beta r_1[Ci2\beta r_2-Ci2\beta r_1]-\sin 2\beta r_1[Si2\beta r_2-Si2\beta r_1]\right\} \tag{33}$$

$$N(l)=30\left\{\sin 2\beta l\left[\frac{r_1}{l}\ln\frac{r_2}{r_1}-1\right]+\cos 2\beta r_1[Si2\beta r_2-Si2\beta r_1]-\sin 2\beta r_1[Ci2\beta r_2-Ci2\beta r_1]\right\}. \tag{34}$$

For the pointed antenna of Fig. 6C, one transformation set may be obtained for the entire region from r_3 to r_1, or the problem may be broken into two cascaded non-uniform lines, separated at the beginning of the taper, $r=r_2$. Calculations were made both ways without much difference in the results, so the results will be given here only for the over-all antenna as a unit. The local characteristic impedance is

$$K(z)=\begin{cases}60\ln\left[\dfrac{2(r_1+z)}{a}\right] & r_1<z<r_2\\[2ex] 60\ln\left[\dfrac{2(r_1+z)l_2}{a(l-z)}\right] & r_2<z<r_3\end{cases}. \tag{35}$$

Substitution of this in Eqs. (28)–(30) yields

$$K_0=60\left[\frac{r_3}{l}\ln\frac{2r_3}{a}-\frac{r_1}{l}\ln\frac{2r_1}{a}-\frac{l_1}{l}\right] \tag{36}$$

$$M(l)=[M(l)]_{cyl}+30\left[\frac{l_2}{l}(1-\cos 2\beta l)+\cos 2\beta l S_1(2\beta l_2)-\sin 2\beta l Si(2\beta l_2)\right] \tag{37}$$

$$N(l)=[N(l)]_{cyl}+30\left\{\sin 2\beta l\left[\frac{l_2}{l}-S_1(2\beta l_2)\right]-\cos 2\beta l Si(2\beta l_2)\right\}, \tag{38}$$

where $[M(l)]_{cyl}$ and $[N(l)]_{cyl}$ denote the M and N for a cylinder extending from r_1 to r_3, computed from Eqs. (33) and (34) with r_3 substituted for r_2. $S_1(x)$ is the function tabulated by Terman,[12] equal to

$$\ln x+0.5772-Ci(x).$$

THE TERMINATING ADMITTANCE

The radiation admittance at the end of the antenna, representing the effect of TM modes inside and outside the antenna region on the TEM mode equivalent circuit, may be taken from Schelkunoff's approximate formulas.[13] For a monopole, they are

$$Y_R=\frac{R_a(\beta r_2)+jX_a(\beta r_2)}{K_0^2}, \tag{39}$$

[11] S. A. Schelkunoff, *Electromagnetic Waves* (D. Van Nostrand Company, Inc., New York, 1943), p. 291.

[12] F. E. Terman, *Radio Engineers Handbook* (McGraw-Hill Book Company, Inc., New York, 1943), p. 17.

[13] See reference 11, pp. 465 and 453.

where

$$R_a(\beta r_2)=15\{2S_1(2\beta r_2)+\cos 2\beta r_2 \times[C+\ln\beta r_2-2Ci2\beta r_2+Ci4\beta r_2] +\sin 2\beta r_2[Si4\beta r_2-2Si2\beta r_2]\} \quad (40)$$

$$X_a(\beta r_2)=15\{2Si(2\beta r_2)+\sin 2\beta r_2 \times[Ci4\beta r_2-\ln\beta r_2-C]-\cos 2\beta r_2 Si4\beta r_2\}. \quad (41)$$

$$C=0.5772.$$

The approximations referred to in the development of the above limit results to relatively thin antennas. Schelkunoff has mentioned a lower limit for K_0 of around 500 ohms (for a dipole antenna) in order for the results to be applicable. The antennas studied in this report are very close to this limit.

There is also a question of a possible added admittance at the end to account for the caps or other end effect.[14] It was believed that the theory of this end effect is insufficiently understood to justify its inclusion in the calculation. The experimental results actually showed no measurable difference between the capped antenna of Fig. 6b and the open-ended antenna of Fig. 6a—a result checking similar experiments by Brown and Woodward.[1] Of course this might mean only that the two antennas had similar end effects, and not that the end effect for either is negligible. The

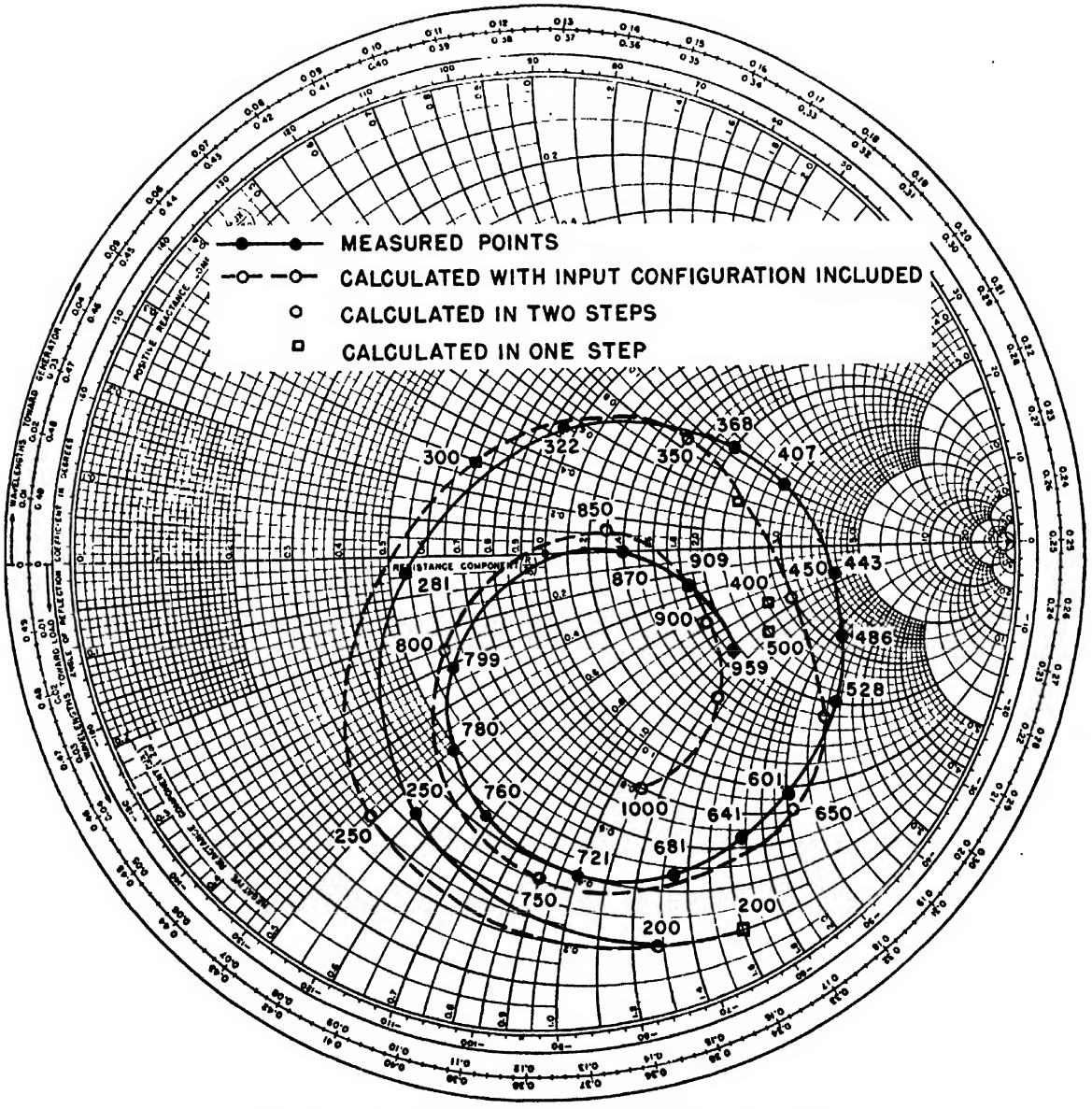


FIG. 15. Input impedance of pointed antenna referred to 66-ohm line.

[14] See reference 11, p. 465.

pointed antenna was chosen to minimize this end effect, but, since it represents a more non-uniform transmission system, the error in that stage of the analysis may be greater so that the net improvement is doubtful.

CALCULATED AND MEASURED IMPEDANCES FOR THE OVER-ALL ANTENNA

All parts of the equivalent circuit of Fig. 5 have now been described so that a calculation of the input impedance referred to the end of the antenna may be accomplished by straightforward, though lengthy, network and transmission line calculations.

The input impedance curves for the antennas sketched in Fig. 6 were measured over the range 200 to 1000 mc/sec. for comparison with the calculated values. The general set-up for the measurement is indicated in Fig. 13. The monopoles were extended horizontally from the side of a building on which there was a vertical ground plane consisting of a three-foot-square copper sheet set into a 20-foot-square section of one-fourth inch galvanized mesh. Each antenna was screwed into the one-fourth inch diameter inner conductor of the 66-ohm coaxial Chipman line.[15] Signal generator power was supplied to a semicircular loop, two inches in diameter, approximately one foot from the antenna. Received power was taken from a small loop (approximately a $\frac{1}{16}$-inch diameter semicircle) in the shorting plunger of the line by cable to an APR receiver with external meter used as a detector. The standing wave ratio (ratio of maximum current to minimum current as the shorting plunger is moved) was found by several methods, all of them utilizing comparison with an auxiliary signal generator having a calibrated attenuator. Position of the maximum current reference was also observed so that each point could be plotted directly on a Smith transmission line chart, and then translated to total impedance or admittance if desired.

Since no measurable difference was observed between the results for the capped and open-ended cylinders, only one comparison curve is shown for the cylindrical antennas. Figure 14 shows the measured curves of resistance and reactance *versus* frequency, compared with calculated values utilizing the methods of this paper, and also with curves calculated without consideration of the input configuration. These latter calculations simply extend the non uniform line to $r=0$, as explained in the discussion of Fig. 2. The comparison of calculated and measured results for the pointed antenna is shown directly in Smith chart form in Fig. 15.

In comparing the calculated and measured curves, it is concluded that the agreement, although not perfect, is greatly improved by the consideration of the input configuration. This is true for both the cylindrical and the pointed antennas. It is believed that much of the remaining disagreement is caused by the relatively large diameter-to-length ratios of the antennas used, which make certain aspects of the thin antenna theory used subject to error. The end effect at the outer end of the antenna is also an unknown and may be important. Of course there are also possible sources of error in the measurement procedure. Among these are the finite extent and conductivity of the ground plane, the presence of reflections from nearby buildings, the finite conductivity of the measurement line, the finite size of the coupling loops, and random instabilities occasionally present in the signal generators.

STUDIES WITH VARIOUS INPUT CONFIGURATIONS

The study described has utilized a given input configuration and different antennas. G. W. Zeoli[16] has recently applied the approach of this paper to the study of given antennas with different input configurations. Two different input configurations were used for a $\frac{1}{2}$-inch-diameter antenna 11 inches long, and two for a $\frac{1}{4}$-inch-diameter antenna 11 inches long. There was a considerable difference between the measured results for a given antenna with the different input configurations, and the amount of this difference was predicted reasonably well by the theoretical calculations. This in some ways gives a more satisfying check of the principles introduced here than the measurements cited in the preceding section. The agreement in Zeoli's results was best for the thin antennas, checking our belief that some of the deviations in Fig. 15 arise from the application of thin antenna theory to antennas which are on the borderline of its applicability.

Work has also been reported in the literature by King and co-workers[17, 18] on antennas fed by parallel-wire lines. A terminal zone is defined there also, and, although the method of attack is somewhat different from that of this paper, equivalent networks are obtained by averaging processes for representation of the terminal zone.

APPENDIX A. A QUASI-STATIC METHOD FOR OBTAINING EQUIVALENT CIRCUITS

It is assumed that the electric field distribution is represented well enough by that obtained from a static field map, and the region to be studied is divided by the flux tube boundaries into equal flux divisions, as indicated in Fig. 8. A capacitance C_n (equal for all subdivisions) and an inductance L_n (computed from Eq. (10) for axially symmetric regions) may be found. If the subdivision is represented by a symmetrical T equivalent circuit, the $\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}$ constants are

$$\mathfrak{A}_n = \left(1 - \frac{\omega^2}{2\omega_n^2}\right)$$

[15] R. A. Chipman, J. App. Phys. 10, 27–38 (1939).

[16] G. W. Zeoli, "Impedance of antennas with various input configurations," Report No. 157, Contract NObsr-39401, Antenna Laboratory, University of California, August 22, 1949.

[17] R. King, J. App. Phys. 20, 832 (1949).

[18] R. King and K. Tomiyasu, Proc. I.R.E. 37, 1134 (1949).

$$\mathfrak{B}_n = j\omega L_n\left(1-\frac{\omega^2}{4\omega_n^2}\right)$$

$$\mathfrak{C}_n = j\omega C_n$$

$$\mathfrak{D}_n = 1-\frac{\omega^2}{2\omega_n^2},$$

where

$$\omega_n^2 = \omega^2 L_n C_n.$$

The N subdivisions are now cascaded and the overall $\mathfrak{ABCD}$ constants obtained by multiplying the individual matrices together. Multiplying from the right, the product of the last two matrices is taken, neglecting powers of (ω/ω_n) higher than the second.

$$\mathfrak{A}_{N-1,N} = 1-\frac{\omega^2 C_n}{2}(L_N+3L_{N-1})$$

$$\mathfrak{B}_{N-1,N} = j\omega(L_N+L_{N-1})$$

$$\mathfrak{C}_{N-1,N} = 2j\omega C_n$$

$$\mathfrak{D}_{N-1,N} = 1-\frac{\omega^2 C_n}{2}(3L_N+L_{N-1}).$$

Repetition of this process with the succeeding adjacent matrices reveals the law of formation, as follows:

$$\left.\begin{aligned}
\mathfrak{A} &= 1-\frac{\omega^2 C_n}{2}\sum_{k=1}^{N} L_k[2(N-k)+1]\\
\mathfrak{B} &= j\omega\sum_{k=1}^{N} L_k\\
\mathfrak{C} &= NC_n\\
\mathfrak{D} &= 1-\frac{\omega^2 C_n}{2}\sum_{k=1}^{N}(2k-1)L_k.
\end{aligned}\right\} \tag{42}$$

The above may be shown to be the constants for a pi-network as shown in Fig. 9 with L, C_a, and C_b given by Eqs. (5) and (7). The same formulas are applicable to a two-dimensional region, except that the formulas for inductance of the subdivisions (10) must be modified.

APPENDIX B. NETWORK FORMULATION OF NON-UNIFORM LINE THEORY

For a slightly non-uniform transmission system with local impedance per unit length given as $Z(z)=Z_0+\hat{Z}(z)$, and the local admittance per unit length given as $Y(z)=Y_0+\hat{Y}(z)$, Schelkunoff[19] has given the zero-order solution in his Eqs. (11-7), and the first-order solution, obtained by a perturbation calculation, in Eqs. (12-1). The zero- and first-order solutions may be added and put in matrix form as follows:

$$\begin{bmatrix} V(0)\\ I(0)\end{bmatrix} = \begin{bmatrix} \mathfrak{A}(l) & \mathfrak{B}(l)\\ \mathfrak{C}(l) & \mathfrak{D}(l)\end{bmatrix}\begin{bmatrix} V(l)\\ I(l)\end{bmatrix}, \tag{43}$$

where

$$\left.\begin{aligned}
\mathfrak{A}(l) &= [1-F_2(l)]\cosh\gamma_0 l + [F_1(l)+F_3(l)]\sinh\gamma_0 l\\
\mathfrak{B}(l) &= K_0\{[1-F_2(l)]\sinh\gamma_0 l + [F_1(l)+F_3(l)]\cosh\gamma_0 l\}\\
\mathfrak{C}(l) &= \frac{1}{K_0}\{[F_3(l)-F_1(l)]\cosh\gamma_0 l + [1+F_2(l)]\sinh\gamma_0 l\}\\
\mathfrak{D}(l) &= [1+F_2(l)]\cosh\gamma_0 l + [F_3(l)-F_1(l)]\sinh\gamma_0 l
\end{aligned}\right\} \tag{44}$$

$$\left.\begin{aligned}
F_1(l) &= \frac{1}{2}\int_0^l\left[\frac{\hat{Z}(z)}{K_0}-K_0\hat{Y}(z)\right]\cosh 2\gamma_0 z\,dz\\
F_2(l) &= \frac{1}{2}\int_0^l\left[\frac{\hat{Z}(z)}{K_0}-K_0\hat{Y}(z)\right]\sinh 2\gamma_0 z\,dz\\
F_3(l) &= \frac{1}{2}\int_0^l\left[\frac{\hat{Z}(z)}{K_0}+K_0\hat{Y}(z)\right]dz.
\end{aligned}\right\} \tag{45}$$

K_0 and γ_0 are the average characteristic impedance and propagation constant, respectively.

The matrix of Eq. (43) defined by (44), may be recognized as the product of two matrices.

$$\begin{bmatrix} \mathfrak{A}(l) & \mathfrak{B}(l)\\ \mathfrak{C}(l) & \mathfrak{D}(l)\end{bmatrix} = \begin{bmatrix} \mathfrak{A}_1(l) & \mathfrak{B}_1(l)\\ \mathfrak{C}_1(l) & \mathfrak{D}_1(l)\end{bmatrix} \times \begin{bmatrix} \cosh\gamma_0 l & K_0\sinh\gamma_0 l\\ \frac{1}{K_0}\sinh\gamma_0 l & \cosh\gamma_0 l\end{bmatrix}, \tag{46}$$

where

$$\left.\begin{aligned}
\mathfrak{A}_1(l) &= 1-F_2(l)\\
\mathfrak{B}_1(l) &= K_0[F_1(l)+F_3(l)]\\
\mathfrak{C}_1(l) &= \frac{1}{K_0}[F_3(l)-F_1(l)]\\
\mathfrak{D}_1(l) &= 1+F_2(l).
\end{aligned}\right\} \tag{47}$$

[19] See reference 11, Sections 7.11 and 7.12.

Thus the over-all transformation network for the non-uniform line may be broken into a section of uniform line of characteristic impedance K_0 and propagation constant γ_0, with a transforming network with constants given by Eqs. (47) in cascade with its input.

The non-uniform systems met with in this study are specialized in that they are loss free, and also have a constant LC product for each point along the line. The following may then be shown:

$$\frac{\hat{Z}}{K_0}+K_0\hat{Y}=0$$

$$\frac{\hat{Z}}{K_0}-K_0\hat{Y}=2j\omega(L_0C_0)^{\frac{1}{2}}\left[\frac{K(z)}{K_0}-1\right].$$

Substitution of these relations in Eqs. (61) gives

$$F_1(l)=\frac{N(l)}{jK_0}$$
$$F_2(l)=\frac{M(l)}{K_0} \qquad (48)$$
$$F_3(l)=0,$$

where

$$M(l)=\beta\int_0^l[K_0-K(z)]\sin 2\beta z dz$$
$$N(l)=\beta\int_0^l[K_0-K(z)]\cos 2\beta z dz. \qquad (49)$$

The $\mathfrak{A}_1$, $\mathfrak{B}_1$, $\mathfrak{C}_1$, $\mathfrak{D}_1$, constants of Eqs. (47) then become as given in the main body of the paper, Eqs. (24)–(27).

Microwave Techniques for the Measurement of the Dielectric Constant of Fibers and Films of High Polymers

T. M. Shaw and J. J. Windle
Western Regional Research Laboratory, Albany, California*
(Received March 27, 1950)

A resonant cavity method is described for the determination of the dielectric constant of polymeric solids in the form of fibers, films, and thin cylinders. Methods for investigating electrical anisotropy are outlined. In the case of fibers, for the electric field parallel to the fiber axis, the dielectric constant is determined directly. For the electric field perpendicular to the fiber axis, the problem arises of calculating the dielectric constant of the fiber from that of an air-fiber mixture. For the latter purpose the Lichtenecker equation was found to be applicable. Illustrative measurements of the dielectric constant of wool, Nylon, and Cellophane are presented and discussed briefly.

INTRODUCTION

THE development of sources of microwave energy has made possible the extension of dielectric-constant studies on solid polymers to frequencies in the region 3000 to 30,000 Mc (1 Mc$=10^6$ cycles per second) or more where new facts concerning the structure of polymers are to be expected. The keratins represent a group of polymers for which dielectric studies in the microwave range should be especially informative. Dielectric constants for wool keratin[1] and for horn keratin[2] have been reported for frequencies between about 500 c.p.s. and 13 Mc. At 13 Mc the dielectric constant was found to be 4.2. This value is considerably greater than the square of the index of refraction for visible light ($\sim$2.25) and has been interpreted as an indication of atomic or ionic polarization.[1] This should lead to a large variation in the dielectric constant with frequency in the microwave region. In order to investigate this possibility, suitable methods were sought for making dielectric constant measurements on the keratins in the microwave region.

Since the keratins of greatest interest are available only in limited forms, for example wool fibers, the generally used cavity resonator[3] and wave-guide procedures,[4,5] which employ relatively large accurately machined specimens, are not directly applicable. For this reason a modified cavity resonator technique has been developed for use with dielectric specimens in the form of fibers, films, and thin cylinders. The procedures and apparatus are described in the present paper. They have been used exclusively in the 3000-Mc frequency region but appear to be suitable for use at higher frequencies with only minor modifications.

* Bureau of Agricultural and Industrial Chemistry, Agricultural Research Administration, United States Department of Agriculture. Report of a study made under the Research and Marketing Act of 1946.

[1] J. Errera and H. S. Sack, Ind. Eng. Chem. 35, 712 (1943)
[2] G. King, Trans. Faraday Soc. 43, 601 (1947).
[3] Horner, Taylor, Dunsmuir, Lamb, and Jackson, J. Inst. Elec. Eng. 93, 53 (1946).
[4] S. Roberts and A. von Hippel, J. App. Phys. 17, 610 (1946).
[5] For a general discussion of methods for making dielectric constant measurements at microwave frequencies see R. M. Redheffer, *Technique of Microwave Measurements* (McGraw-Hill Book Company, Inc., New York, 1947), p. 561.

THEORY

Figure 1 represents a cavity resonator. The cavity is circular in cross section and is closed with plane parallel ends perpendicular to the axis of the cylinder. As is well known, when a cavity is excited to resonance at the frequency f_0, in the fundamental or TM_{010} mode, the electric field in the cavity is directed parallel to the cylinder axis. The electric field possesses a single maximum at the axis and falls to essentially zero at the cylinder wall.

The usefulness of such a resonator for dielectric constant measurements derives from the fact that when a dielectric specimen is placed in the cavity the resonance frequency is changed; such change in frequency can be related quantitatively to the dielectric properties of the specimen.

The theory of the use of the resonator for the case of a smooth circular cylinder low loss[6] dielectric specimen placed along the cavity axis is well known.[3] For a specimen so small in diameter that insertion of the specimen in the cavity results in a negligible change in the electric and magnetic field distribution, the resonant frequency is given by the equation,

$$f=f_0-(\epsilon-1)f_0b^2/0.539a^2. \qquad (1)$$

In this equation f is the frequency of the resonator containing the dielectric specimen, f_0 is the frequency of the empty resonator, ϵ is the dielectric constant of the specimen, b is the radius of the dielectric cylinder, and a is the radius of the cavity resonator.

The frequency change (f_0-f) to be expected when a single wool fiber of a length equal to that of the cavity is used as a dielectric specimen is easily obtained by means of Eq. (1). Assuming as typical values, $\epsilon\sim4$, $b\sim10^{-3}$ cm, $a=4.5$ cm, $f_0=3000$ Mc, it is found that $(f_0-f)\sim0.001$ Mc. Although this small frequency change can be detected readily with standard microwave equipment, it is too small for accurate measurement purposes. Since changes in the temperature of the cavity and in the humidity of the air in the cavity can produce relatively large frequency changes it is necessary to work with specimens consisting of several thousand fibers. In this way only moderately constant temperature and humidity are required.[7]

A cylindrical bundle containing several thousand wool fibers is about 0.15 cm in radius. The theoretical calculations of Horner and co-workers[3] show that if Eq. (1) is applied to a cylinder of dielectric of about this size and with a dielectric constant of six or less, the frequency change will be accurate to within one percent. This is true because for a cylinder of dielectric of this diameter the variation of the intensity of the electric field across the cylinder cross section is negligible.

For convenience in calculations Eq. (1) may be expressed in terms of the volume of the dielectric specimen and the volume of the cavity resonator by replacing b^2 by its equivalent $v_s/\pi l$ and a^2 by its equivalent $v_c/\pi l$. The resultant equation is

$$f=f_0-(\epsilon-1)f_0v_s/0.539v_c. \qquad (1a)$$

In this equation it is understood of course that the volume v_s is that of a specimen extending the entire length of the cavity.

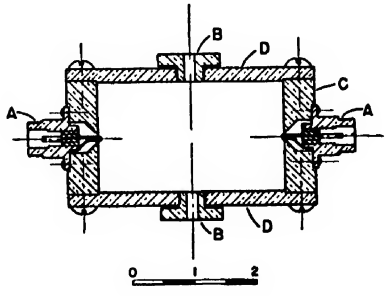


FIG. 1. Cross-sectional view of cylindrical cavity resonator. *A*, Type *N* cable connectors; *B*, plugs to permit specimen to be inserted along axis of resonator; *C*, cylinder wall, *D*, end plates perpendicular to cylinder axis.

Equation (1) and thus also (1a) was derived for a dielectric specimen consisting of a single circular cylinder. Since a bundle of fibers only approximates a circular cylinder in shape, tests were made to determine whether errors were introduced by applying Eq. (1a) to a specimen composed of a number of circular cylinders. Glass was chosen as a convenient test material rather than a polymer since it was desired to avoid possible effects due to anisotropy and to water absorption. Single glass cylindrical rods were prepared with radii ranging from about 0.02 to 0.05 cm. These rods were inserted in the cavity individually and in groups and the frequency change (f_0-f) determined. The results are shown in Table I. Constancy of the ratio $(f_0-f)/b^2$ shows that the frequency change (f_0-f) is a linear function of the volume of glass contained in a specimen irrespective of whether the glass is in the form of a single rod or a group of rods. Calculations showed that the external radius of the largest bundle of rods considered was within the limit for the size of a single cylinder set by Horner[3] if an accuracy of one percent or better is to be expected. Table I also shows that dielectric constants calculated from the various measurements are in good agreement. The statistical error (standard deviation) is slightly less than one percent which, considering the approximations made in the derivation, is all that can be expected from Eq. (1a) for specimens of this size and dielectric constant.

These results demonstrate that the exact form of the

[6] The equations given below apply strictly only to a hypothetical dielectric with a zero loss tangent, i.e., $\tan\delta=0$. Their application to the materials considered in the present paper is justified because $\tan\delta$ does not exceed 0.05 and only a negligible error is introduced. For a complete discussion see reference 5, p. 568.

[7] For the particular 3000-Mc cavity resonator used a one-degree rise in temperature produces a decrease in the resonant frequency of 0.058 Mc. A change in relative humidity from 25 percent to 50 percent will cause a frequency decrease of about 0.03 Mc.

boundary of the dielectric specimen is not important provided the specimen is confined to a region where the electric field is essentially constant as discussed above. Since the shape of the specimen is not important, it was assumed and subsequently verified with Teflon films and rods that the dielectric specimen may have any of the forms *A*, *B*, *C*, *E*, or *F* of Fig. 2 provided only that the specimen is continuous over the entire length of the cavity resonator and be composed of elements of solid the boundaries of which are tangential to the electric field. For all such specimens Eq. (1a) can be used to calculate the dielectric constant. The volume of the specimen is conveniently and accurately determined from its mass and density. This procedure is especially convenient for fiber specimens where the small size and large number of the fibers in a specimen makes accurate measurements of the fiber volume very difficult.

TABLE I. Summary of measurements of dielectric constant of Pyrex glass (No. 7740) rods.

Specimen numbers	Square of radius of rod b^2	Change in frequency (f_0-f)	$(f_0-f)/b^2$	Dielectric constant
	$(cm)^2$	Mc		
1	27.10×10^{-4}	3.82	1.410×10^3	4.66
2	26.80	3.83	1.430	4.70
3	22.40	3.18	1.420	4.68
4	21.80	3.06	1.405	4.64
5	19.05	2.66	1.395	4.62
6	15.95	2.24	1.405	4.64
7	12.52	1.79	1.430	4.70
8	9.36	1.32	1.410	4.65
9	9.12	1.32	1.448	4.73
10	6.76	0.972	1.439	4.73
11	4.68	0.686	1.464	4.79
1, 3, 4	71.30	10.15	1.425	4.68
6, 7, 8	37.83	5.41	1.430	4.70
9, 10, 11	20.56	2.96	1.440	4.73
2, 7, 8, 11	53.36	7.61	1.428	4.69
		Average	1.425 ± 0.018	4.69 ± 0.04

Specimens of type *E* are of interest when it is desired to investigate the electrical anisotropy in polymer films.[8] The general treatment of *E* has been discussed. It remains only to remark that by rolling the film into a thin cylinder with the selected direction in the film parallel to the axis of the cylinder it is possible to determine the dielectric constant of the film with the electric field parallel to the direction chosen. By rolling the film into cylinders with the axis in other directions anisotropy can be investigated.

The type of specimen represented by *D* of Fig. 2 is perhaps the most important that must be considered. A solenoidal specimen of fibers of this type is used to determine the dielectric constant of the fibers with the electric field perpendicular to the fiber axis. Such measurements when combined with those on parallel bundles of fibers may be used to investigate the electrical anisotropy of fibers.[8]

A solenoidal specimen is obtained by winding fibers in the form of a helix with very small pitch about a central core. The core is made from a material of low dielectric constant and dielectric loss in order to minimize the correction to the frequency change to be made for the core. Polystyrene is satisfactory from this viewpoint but it is not satisfactory when it is necessary to heat the fibers such as is required in some instances to remove absorbed water. It was because of the latter limitation that quartz was usually used in the present work.

The dielectric constant which results when Eq. (1a) is applied to a solenoidal specimen is that of an *air-solid* mixture. The volume of the specimen v_s used in Eq. (1a) is calculated from the external diameters of the solenoid and its core and the length of the solenoid. The length is, of course, the same as that of the resonator. From these measurements the dielectric constant of the air-solid mixture (fibers plus air) is calculated. The justification for this procedure lies in the fact that the mixture of fibers and air occupies an essentially uniform electric field and the frequency changes due to the core and to the fiber-air mixture are additive.

To deduce the dielectric constant of the solid material (fibers) from that of the air-fiber mixture recourse must be had to a mixture law. For this purpose consideration has been given to only the equations of Lorentz[9] and of Lichtenecker,[10] for dielectric mixtures. For a two-component system the Lorentz equation is

$$(\epsilon_m-1)/(\epsilon_m+2)=v_f(\epsilon_\perp-1)/(\epsilon_\perp+2)+v_a(\epsilon_a-1)/(\epsilon_a+2). \quad (2)$$

In this equation ϵ_m is the dielectric constant of the air-fiber mixture, $\epsilon_\perp$ is the dielectric constant of the fiber substance measured with the electric field directed perpendicular to the fiber axis, ϵ_a is the dielectric constant of air and v_f and v_a are, respectively, the volume of the fiber substance and of air in 1 cc of the fiber-air mixture.

Similarly, for a two-component mixture the Lichtenecker equation may be written,

$$\log\epsilon_m=v_f\log\epsilon_\perp+v_a\log\epsilon_a, \quad (3)$$

and the symbols have the same meanings stated for Eq. (2).

Owing to the well-recognized theoretical limitations of mixture equations for dielectrics it was desired to ascertain only on a purely empirical basis whether one of these equations would give satisfactory values for $\epsilon_\perp$. The validity of the equations was tested by comparing the value of $\epsilon_{||}$ determined for glass fibers with $\epsilon_\perp$ for the same fibers as calculated by means of the mixture equa-

[8] A preliminary discussion of methods for measurement of electrical anisotropy in polymers at microwave frequencies was presented by the authors at the Seattle meeting of the American Physical Society, June 27–29, 1949; T. M. Shaw and J. J. Windle, Phys. Rev. 76, 586 (1949).

[9] H. A. Lorentz, *Theory of Electrons* (Teubner, Berlin, 1916), second edition, p. 147.

[10] K. Lichtenecker, Physik. Zeits. 27, 115 (1926).

tions from measurements performed on solenoidal specimens. For glass fibers $\epsilon_{\parallel}$ should equal $\epsilon_{\perp}$ since well-annealed glass is isotropic. Table II summarizes the experimental data and calculations for a number of specimens of glass fibers. The table shows that the mean value of $\epsilon_{\perp}$ calculated with the aid of the Lichtenecker equation agrees with $\epsilon_{\parallel}$ within about two percent. This is considered to be satisfactory because of the likelihood of errors which can arise in the measurement of the external diameter of the solenoid. An error of one percent in diameter of the solenoid can cause an error of about two percent in $\epsilon_{\perp}$. It is significant in this regard that a low value of $\epsilon_{\perp}$ is to be associated with a low value in the diameter of the solenoid. This is the most likely direction for error since it is not possible to define exactly the outer boundary of the air-fiber mixture.

The values of $\epsilon_{\perp}$ obtained by means of Eq. (3) are higher than $\epsilon_{\parallel}$ by about 7 percent. Because the disagreement between $\epsilon_{\parallel}$ and $\epsilon_{\perp}$ obtained with the Lichtenecker equation is no greater than the experimental error it was adopted as satisfactory for calculating the dielectric constant of the fibers from measurements on an air-fiber mixture in the form of a solenoid. The table also shows that ϵ_m for the air-fiber mixture as well as $\epsilon_{\perp}$ is independent of the size of the solenoidal specimen. This constitutes further experimental proof that the electric field is essentially uniform over the cross section occupied by the solenoidal dielectric specimens.

EXPERIMENTAL

A cross-sectional view of the cavity resonator is shown in Fig. 1. The cavity dimensions are almost identical with those for a similar cavity described by Horner.[3] The present cavity was equipped with standard type *N* coaxial cable connectors which terminate in small coupling loops. These loops are inserted into the cavity through holes in the cylinder wall. The plane ends of the cavity are equipped with removable plugs in order to insert dielectric specimens. For non-hygroscopic specimens the removable plugs are simply drilled with holes large enough to accommodate the specimen. The effect of the holes and the specimen contained within them on the resonant frequency of the cavity is very small.[11] For hygroscopic specimens the removable plugs are designed to hold the specimen in place in the cavity and to allow the cavity to be evacuated.

When the cavity was used at temperatures other than room temperature a correction was made for the change in resonant frequency caused by expansion or contraction of the cavity. The correction required was determined experimentally to be 0.058 Mc per degree C by measuring the resonant frequency of the cavity over the temperature range 0 to 100°C.

A block diagram of the microwave equipment used to determine the frequency of the resonant cavity is shown in Fig. 3. The klystron oscillator is frequency modulated

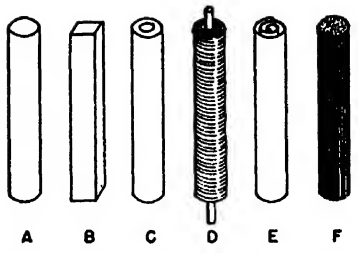


FIG. 2. Dielectric specimens. *A*, solid circular cylinder; *B*, solid rectangular bar; *C*, hollow circular cylinder; *D*, fiber solenoid; *E*, thin film wound into circular cylinder; *F*, bundle of fibers.

by means of a saw-tooth wave derived from the oscilloscope sweep. The signal from the klystron is attenuated and fed simultaneously to the cavity resonator and to a wave meter. The rectified output of the cavity is combined with the rectified output from the wave meter in a balanced mixer transformer. The resultant signal is amplified and displayed on the oscilloscope. The frequency of the oscillator is adjusted until it is being swept over the frequency of the cavity resonator. When this is done, a typical resonance curve is obtained. The wave meter is adjusted until it is tuned to the same frequency as the cavity as indicated by the appearance of a double-humped pattern on the oscilloscope. By proper adjustment of the wave-meter output the pattern obtained on the oscilloscope is a very sensitive indication of the equality of frequency of the cavity and wave meter. The general usefulness of such patterns in the tuning of resonators has been discussed.[12] The same procedure is followed to determine the frequency f_0 of the empty cavity resonator and the frequency, f, of the cavity resonator when it contains a dielectric specimen.

The wave meter used to determine the frequency difference (f_0-f) is especially designed to facilitate the determination of small frequency differences. The wave meter consists of a section of rectangular wave guide tuned by means of movable plungers in the ends. Operation of the wave meter in the TE_{107} mode results in a relatively large displacement in plunger position for

TABLE II. Application of Lichtenecker and Lorentz equations to a mixture of glass fibers and air.

$\epsilon_{\parallel}$ (glass fibers) $=5.99\pm0.05$

Specimen number	Volume of fiber-air mixture (cc)	Volume of fibers (cc)	Dielectric constant of mixture ϵ_m	$\epsilon_{\perp}$ calculated Lichtenecker equation	$\epsilon_{\perp}$ calculated Lorentz equation
1	14.67×10^{-3}	7.10×10^{-3}	2.35	5.86	6.42
2	20.20	9.15	2.24	5.87	6.52
3	35.15	16.10	2.25	5.88	6.39
4	14.92	6.60	2.20	5.98	6.43
5	10.20	4.83	2.31	5.86	6.44
6	17.68	8.51	2.34	5.85	6.42
7	23.88	11.75	2.40	5.93	6.49
			Average	5.89 ± 0.05	6.44 ± 0.05

[11] J. C. Slater, Rev. Mod. Phys. 18, 441 (1946).

[12] R. L. Sproull and E. G. Linder, Proc. I.R.E. 34, 305 (1946).

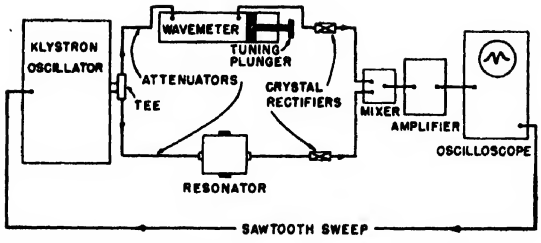


FIG. 3. Block diagram of microwave apparatus.

a small frequency difference (f_0-f). The calibration for the wave meter was calculated from the dimensions of the wave guide. Rotation of the screw used to drive the tuning plunger by one turn changes the frequency by 2.756 Mc. The sensitivity of the measuring system is such that a change in frequency of 0.003 Mc can be detected easily.

Verification of the calculated calibration of the wave meter was accomplished indirectly by measuring the dielectric constants of a number of materials for which the dielectric constants at 3000 Mc are given in the literature.[13] Figure 4 is a graph comparing the values obtained by the present method with the values taken from the literature.[13] With the exception of the data for glasses, the values given in the literature and those determined by the methods described in this paper agree to within about one percent. For the glasses deviations up to about five percent were obtained. In view of the discrepancy in the low frequency dielectric constants for glasses reported by different sources, it was assumed that the deviations greater than one percent observed in the microwave dielectric constants are due to differences in composition of the specimens rather than to errors in measurement. This assumption is supported by the

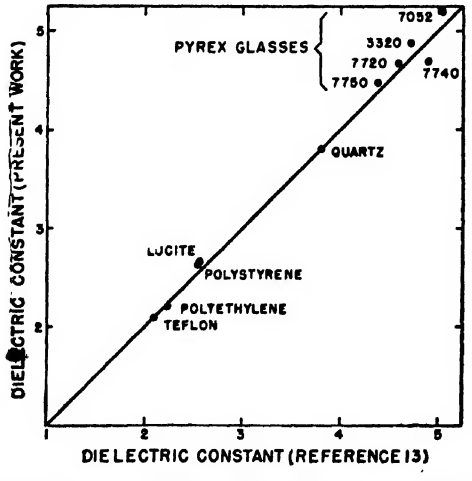


FIG. 4. Comparison between dielectric constants measured by methods described in present paper and data of A. von Hippel (reference 13).

[13] A. Von Hippel, Tables of Dielectric Materials, Vols. I and II, Laboratory for Insulation Research, M.I.T. Cambridge, Massachusetts, NDRC Division 14, Report No. 266 and 282.

measurements summarized in Table I which show that variations in the dielectric constant for various sizes of samples of a given glass are much smaller than the discrepancy between the literature values and the present determinations. For the data in Table I the standard deviation is less than one percent.

TYPICAL RESULTS

The dielectric constants of a number of natural and synthetic polymers have been measured by the methods outlined in this paper. The measurements presented below have been selected to illustrate a number of the types of specimen which have been studied.

The dielectric constant of dry wool fibers has been determined for the electric field parallel to the fiber axis ($\epsilon_{\parallel}$) and for the electric field perpendicular to the fiber axis ($\epsilon_{\perp}$). The fibers were arranged into bundles and solenoids as illustrated in Fig. 2 and dried at 100°C in a vacuum oven. The results are shown in Table III. The values of the dielectric constant for the individual specimens agree to within about plus or minus 3 percent. This variation is believed to be due to errors in the dimensions of the specimens. The wool fibers are not uniform and it is difficult to prepare specimens that are well defined geometrically. If methods can be perfected to prepare more uniform specimens, the variation in the dielectric constant can probably be reduced substantially.

The differences between the values of $\epsilon_{\parallel}$ and $\epsilon_{\perp}$ for wool shown in Table III are sufficiently great compared to the statistical errors that it appears likely wool is electrically anisotropic at 3000 Mc. This is interesting in view of the anisotropy reported for another keratin[2] (horn) at much lower frequencies and in view of the apparent lack of anisotropy in horn at 3000 Mc.[8] The validity of the conclusion that wool is electrically anisotropic is of course based on the applicability of the

TABLE III. Dielectric constant of dry wool at 3000 Mc.

Specimen number	Dielectric constant $\epsilon_{\parallel}$*
1	4.01
2	3.66
3	3.86
4	3.75
5	4.03
6	3.97
	Average 3.88±0.15
Specimen number	**Dielectric constant $\epsilon_{\perp}$***
7	4.40
8	4.31
9	4.38
10	4.56
	Average 4.41±0.11

* $\epsilon_{\parallel}$ = dielectric constant for electric field parallel to fiber axis. $\epsilon_{\perp}$ = dielectric constant for electric field perpendicular to fiber axis.

TABLE IV. Dielectric constant of Nylon fibers at 3000 Mc (fibers stretched about 300 percent before measurements, moisture content 2.0±0.5 percent).

Specimen number	Dielectric constant ϵ_{11}*
1	3.18
2	3.14
3	3.09
4	3.11
	Average 3.13±0.04

* ϵ_{11} = dielectric constant for electric field parallel to fiber axis.

TABLE V. Dielectric constant of dry Cellophane film (electric field parallel to plane of film).

Specimen	Dielectric constant
1	4.05
2	4.06
3	4.04
4	4.03
5	4.00
6	4.06
7	4.01
8	4.06
9	4.10
10	4.02
	Average 4.04±0.03

Lichtenecker equations to mixtures of fibers and air as discussed earlier. If these equations are not valid for wool, then the conclusion of an anisotropy for wool would have to be modified. Further discussion of the anisotropy in wool and other related polymers will be reserved for discussion at a later time.

Table IV contains values of the dielectric constant (ϵ_{11}) of Nylon fibers. The agreement between the individual measurements of (ϵ_{11}) is better than was obtained for wool. This is probably a result of the fact that the Nylon fibers measured were in the form of a uniform continuous filament. This made it possible to prepare more uniform specimens than was possible with wool.

In Table V results are presented which illustrate an application to measurement of the dielectric constant of a polymer film. The material in this case is Cellophane. The ten specimens for which data are given were cut from a single sheet and rolled to form cylinders in the manner illustrated by E of Fig. 2. The individual values of the dielectric constant agree to within about plus or minus one percent, which indicates a high degree of uniformity in a single sheet of film. Similar measurements on specimens from a second sheet of film gave an average value for the dielectric constant of 3.89 compared to the value of 4.04 obtained with the first sheet. This difference in the dielectric constant of the two sheets probably represents a real difference in the Cellophane.

From the results presented it may be concluded that the resonant cavity method may be used to advantage for the determination of the dielectric constants of polymers either in the form of thin cylinders, fibers, or films. It thus appears to be especially useful in the study of the dielectric properties of natural polymers most of which are available only in highly specialized or limited forms such as, for example, fine fibers. The method appears to be capable of yielding accurate values of the dielectric constant even when used with rather small amounts of material.

A more complete discussion of the dielectric constants of wool and related polymers obtained by the methods described above will be given in a subsequent paper to be submitted to this journal.

Otto A. Beeck

It is with sincere regret that we announce the death on July 5, 1950 of Dr. Otto A. Beeck, Associate Editor of the *Journal of Applied Physics*.

Determination of Residual Stresses of Quenching Origin in Solid and Concentric Hollow Cylinders from Interferometric Observations

A. W. Sáenz
Applied Mathematics Branch, Mechanics Division, Naval Research Laboratory, Washington, D. C.
(Received August 10, 1949)

A theoretical procedure for the determination of the stresses of quenching origin in solid and concentric hollow cylinders has been derived. The method relies on interferometric data and the assumptions discussed in Section I, whose validity for the technically important case of glass is supported by a number of experimental investigations.

The results should be of interest for the analysis of stresses in quenched glass cylinders, so great in magnitude that the usual cutting techniques are precluded, because of the explosive characteristics of highly stressed objects of this material.

I. INTRODUCTORY REMARKS

An optical method for the determination of the pointwise distribution of stresses of thermoelastic or residual type due to quenching in transparent initially isotropic spheres and infinite cylinders was developed recently at the University of Michigan, and presented in an earlier paper.[1] In the former case, it was proved that the residual stresses could be measured by means of a Babinet compensator, provided the photoelastic law was valid for these stresses, while in the latter it was necessary to introduce special assumptions concerning the mechanism of quenching. To avoid the use of such hypotheses, a theoretical method for the determination of residual stresses of quenching origin in solid and concentric hollow cylinders from interferometric observations alone will be developed in this article. This extension of the use of the usual interferometric procedures[2] to three-dimensional stress fields will complete the work on spheres and cylinders mentioned previously. The method should be of particular interest for the case of glass.

The considerations in the following pages rest on three assumptions: (a) The directions of the principal stresses are assumed to be parallel to the cylindrical coordinates ρ, φ, and z. (b) The stress system resulting from quenching is assumed to be radially symmetrical, i.e., it is independent of φ and z. (c) The Maxwell-Neumann stress-optic law holds, i.e.,

$$\begin{aligned} n_P - n_0 &= C_1 P + C_2(Q+R), \\ n_Q - n_0 &= C_1 Q + C_2(P+R), \end{aligned} \qquad (1)$$

where P and Q are secondary principal stresses in the plane of the wave front, R is the stress normal to the wave front, n_P and n_Q are refractive indices for plane polarized waves with directions of vibration parallel to P and Q, respectively. As to the meaning of n_0 in Eq. (1), a few words are necessary at this point. In the case of stresses below the elastic limit induced by mechanical, thermal, or other means, the initial index of refraction of the material involved is equal to the final index, when the external forces are removed. This is the ordinary case treated in photoelasticity, and experiment shows that (1) is valid if n_0 is identified with the initial refractive index of the substance under examination. However, when residual stresses are being studied photoelastically, the index of refraction of the material is frequently history-dependent, the case of glass furnishing a good illustration of this statement.[3] To avoid these difficulties, the stress-optic law (1) will be assumed in this paper with the proviso that n_0 is now a constant on the same footing as C_1 and C_2. An experimental method to find this constant will be given in Section III.

How valid are the assumptions (a)–(c) for infinite cylinders which have been symmetrically quenched? It is clear that (a) and (b) hold for objects of this shape under the quenching conditions mentioned previously, from symmetry arguments. As to the third assumption, it is realized that this is a highly controversial subject at the present time. Notwithstanding, a number of experimental results strongly support the validity of the photoelastic law for residual stresses of quenching origin in glass, for the case when the initial temperature is low enough so that plastic deformations can be neglected. For example, the conditions of static equilibrium are obeyed by stresses in glass, calculated by photoelastic means,[4] and it is also known that the birefringence of quenched objects of this material is affected by cutting and goes to zero at the surface of the cut, which is a good indication that the optical effect can be expressed as a function of the stress alone. Of course, it will be necessary to carry out a more exhaustive experimental investigation before the above assertions can be generally accepted.

[1] R. C. O'Rourke and A. W. Sáenz, Quart. App. Math. (to be published). The distinction between thermoelastic and residual stresses due to quenching, which is made in that article and the present one, is that the former are transient, because of the fact that the initial temperature is below the softening temperature, while the latter are permanent in character and are produced by rapid cooling from temperatures above the softening point.

[2] See, for example, H. Favre, Rev. d'optique 8, 193–213, 241–261, 289–307 (1929).

[3] Tool, Tilton, and Saunders, J. Research Nat. Bur. Stand. 38, 519 (1947).

[4] F. L. Everett and K. A. Parsons, "Progress reports on stresses in insulators," Eng. Research Inst., University of Michigan (January to June, 1948).

II. DETERMINATION OF σ_ρ, σ_φ, AND σ_z IN INFINITE SOLID AND CONCENTRIC HOLLOW CYLINDERS FROM INTERFEROMETRIC DATA

If a symmetrically quenched infinite cylinder is immersed in a liquid whose index of refraction matches its average index and a plane polarized beam of light is

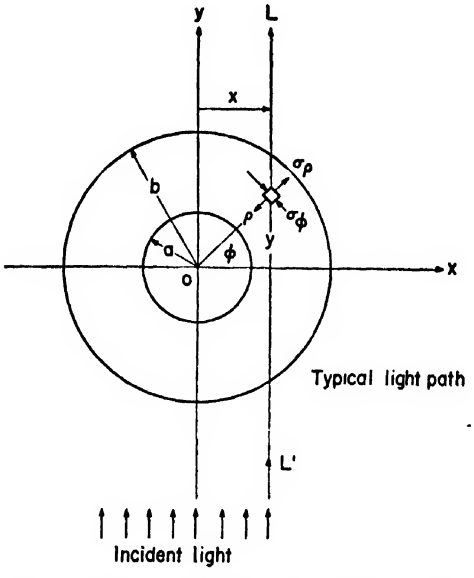


FIG. 1(a) Light normally incident on cylinder (z axis perpendicular to plane of paper).

made to fall normally on it, say along the y axis, as shown in Fig. 1(a), assumption (a) implies:

$$\left.\begin{aligned}\sigma_x&=\sigma_\rho\cos^2\varphi+\sigma_\varphi\sin^2\varphi\\ \sigma_y&=\sigma_\rho\sin^2\varphi+\sigma_\varphi\cos^2\varphi\\ \tau_{xz}&=0\\ \sigma_z&=\sigma_z\end{aligned}\right\}, \tag{2}$$

which means that the following identification can be made:

$$P=\sigma_z,\quad Q=\sigma_x,\quad R=\sigma_y \tag{3}$$

(nothing new would result by putting, instead of (3), $P=\sigma_x$, $Q=\sigma_z$) and, therefore, the directions of the principal stresses do not rotate in the plane of the wave front, so that the usual methods of calculating the phase retardations are correct in this case.

The absolute phase retardations R_P and R_Q along P and Q are, from (1), (2), and (3),

$$\left.\begin{aligned}R_P(x)&=\int_{-[b^2-x^2]^{\frac{1}{2}}}^{[b^2-x^2]^{\frac{1}{2}}}n_P\,dy\\ &=2\int_x^b\{C_1\sigma_z+C_2(\sigma_\rho+\sigma_\varphi)\}\frac{\rho\,d\rho}{[\rho^2-x^2]^{\frac{1}{2}}}\\ &\quad+2n_0[b^2-x^2]^{\frac{1}{2}}\end{aligned}\right\}, \tag{4a}$$

$$\left.\begin{aligned}R_Q(x)&=\int_{-[b^2-x^2]^{\frac{1}{2}}}^{[b^2-x^2]^{\frac{1}{2}}}n_Q\,dy\\ &=2\int_x^b\{C_1(\sigma_\rho\cos^2\varphi+\sigma_\varphi\sin^2\varphi)\\ &\quad+C_2(\sigma_z+\sigma_\rho\sin^2\varphi+\sigma_\varphi\cos^2\varphi)\}\frac{\rho\,d\rho}{[\rho^2-x^2]^{\frac{1}{2}}}\\ &\quad+2n_0[b^2-x^2]^{\frac{1}{2}}\end{aligned}\right\}, \tag{4b}$$

where x is the distance of a given ray LL' [Fig. 1(a)] from the origin 0, measured along the x axis. Equations (4) hold in the interval

$$a\leqq x\leqq b,\quad\text{with}\quad 0\leqq a<b, \tag{4'}$$

where a is the inner radius of the cylinder, taken equal to zero in the case of solid cylinders. Thus, there is no need to distinguish formally between solid and concentric hollow cylinders.

For the sake of completeness, a description of the measurement of $R_P(x)$ and $R_Q(x)$ will be sketched at this point, although it has no essential experimental novelty. A Mach-Zehnder interferometer, as modified by Favre[5] for photoelastic purposes, may be employed

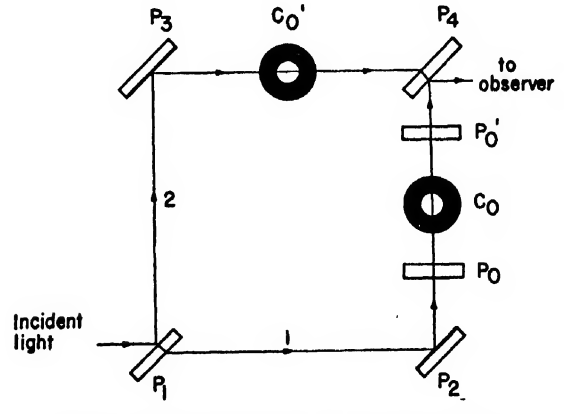


FIG. 1(b) Illustration of measurement of $R_P(x)$ and $R_Q(x)$ by a Mach-Zehnder interferometer.

[Fig. 1(b)]. The plates P_2 and P_3 are full-silvered mirrors, while P_1 and P_2 are half-silvered. The cylinder whose residual stresses are to be determined is placed at C_0, between the half-wave plates P_0 and P_0', in the path of ray 1, and a compensating cylinder of the same size and material is put at C_0', in the path of ray 2. The incident light is plane polarized and has its direction of vibration perpendicular to the plane of the paper [Fig. 1(b)], and the plates P_0 and P_0' permit one to rotate the plane of vibration so that it is parallel to either P or Q, i.e., to the z and x axes, respectively, at C_0

[5] See reference 2, pp. 245–253, where a careful description of the measurement of the absolute retardations along the P and Q axes is given.

and emerges vertical to the plane of the paper after having traversed P_0'. Thus the absolute retardations $R_P(x)$ and $R_Q(x)$ may be measured by Favre's method regardless of the fact that one is dealing with a three-dimensional stress system, because the secondary principal axes do not rotate along the light path for normal incidence.

Regarding $R_P(x)$ and $R_Q(x)$ as known for the interval (4'), it will be demonstrated that σ_ρ, σ_φ, and σ_z can be uniquely determined for all points of the cylinder examined, by inversion of the integral equations (4).

Subtracting (4b) from (4a) and putting $\cos\varphi = x/\rho$, one obtains:

$$R_P(x)-R_Q(x)\equiv r(x)$$

$$=2C\int_x^b\left\{(\sigma_z-\sigma_\varphi)+(\sigma_\varphi-\sigma_\rho)\frac{x^2}{\rho^2}\right\}\frac{\rho d\rho}{[\rho^2-x^2]^{\frac{1}{2}}} \quad (5)$$

where $C=C_1-C_2$ is the relative photoelastic constant and $r(x)$ is the relative retardation at the point x [Fig. 1(a)].

One can eliminate σ_φ from (5) by employing the equation of equilibrium

$$(d\sigma_\rho/d\rho)+[(\sigma_\rho-\sigma_\varphi)/\rho]=0. \quad (6)$$

If one carries out this substitution, integrates (5) partially to get rid of the term in $d\sigma_\rho/d\rho$, and demands that σ_ρ vanish at $\rho=b$, on account of the free boundary, one obtains:

$$r(x)=2C\int_x^b\frac{\sigma_z(\rho)\rho d\rho}{[\rho^2-x^2]^{\frac{1}{2}}}. \quad (7)$$

This is an integral equation of Abel type, index $\frac{1}{2}$,[6] whose only continuous solution is given by

$$\rho\sigma_z(\rho)=-\frac{1}{\pi C}\frac{d}{d\rho}\int_\rho^b\frac{r(x)xdx}{[x^2-\rho^2]^{\frac{1}{2}}},$$

which presents difficulties at $\rho=0$. To arrive at a more convenient expression for σ_z one integrates the last equation by parts[7] and uses the fact that $r(b)=0$, as can be seen from (5), so that the solution of the integral equation (7) becomes

$$\sigma_z(\rho)=-\frac{1}{\pi C}\int_\rho^b\frac{[dr(x)/dx]}{[x^2-\rho^2]^{\frac{1}{2}}}dx. \quad (8)$$

Digressing for a moment at this point, it can be seen that (7) expresses the remarkable fact that, as far as measurements of relative retardation for normal incidence of plane polarized light are concerned, a symmetrically quenched infinite cylinder acts as if only the axial stress σ_z were present. An interesting by-product of (7) and (8) is that σ_z can be determined by the Babinet compensator method, without recourse to more elaborate interferometric procedures.

Returning to the main discussion, one can also find similar expressions for σ_ρ and σ_φ in terms of optical data. With this purpose in mind, (4a) can be written as follows, using (7):

$$\left.\begin{aligned}\int_x^b\{\sigma_\rho(e)+\sigma_\varphi(e)\}\frac{\rho d\rho}{[\rho^2-x^2]^{\frac{1}{2}}}\\ =\frac{1}{2C_2}\{R_p(x)-2n_0[b^2-x^2]^{\frac{1}{2}}\\ -(C_1/C_2)r(x)\}\equiv F(x)\end{aligned}\right\}, \quad (9)$$

where $F(x)$ is known from interferometric observations.

Since (9) is again an Abel integral equation of index $\frac{1}{2}$, one can invert it uniquely in the same way as was done for (7), with the result

$$\sigma_\rho(\rho)+\sigma_\varphi(\rho)=-\frac{2}{\pi}\int_\rho^b\frac{dF(x)/dx}{[x^2-\rho^2]^{\frac{1}{2}}}. \quad (10)$$

From (6) and (10) one has, by repeated application of Dirichlet's formula for the inversion of order in repeated integrals,

$$\left.\begin{aligned}\sigma_\rho(\rho)=-\frac{2}{\pi}\int_\rho^b\frac{S(x)dx}{x^2[x^2-\rho^2]^{\frac{1}{2}}}\\ \text{where}\\ S(x)\equiv\int_x^b dF(x')/dx'\end{aligned}\right\}. \quad (11)$$

Finally, from (6) and (11) it follows that

$$\sigma_\varphi(\rho)=-\frac{2}{\pi}\frac{d}{d\rho}\left\{\rho\int_\rho^b\frac{S(x)dx}{x^2[x^2-\rho^2]^{\frac{1}{2}}}\right\}. \quad (12)$$

Equations (8), (11), and (12) furnish the principal stresses for solid and concentric hollow cylinders once n_0 has been measured. The integrations can be carried out by numerical or analytical methods, but it is felt that the latter are preferable for the most accurate results.[8] Since current interferometric methods in photoelasticity are capable of measuring absolute retardations down to a few percent of a wave-length, the present method seems quite promising with respect to accuracy. It is

[6] E. T. Whittaker and G. N. Watson, *Modern Analysis* (The Cambridge University Press, New York, 1945), p. 229.

[7] This and the succeeding partial integrations and inversions of the order of integration of repeated integrals are justified in virtue of the fact that the absolute retardations, $R_P(x)$ and $R_Q(x)$, as shown by physical considerations, are continuous in the closed interval (4') and possess derivatives which are also continuous in this interval, except at $x=b$, where they have singularities $O\{(b^2-x^2)^{-\frac{1}{2}}\}$.

[8] Analytical procedures were employed in conjunction with the fitting of the retardation curves by polynomials, using least squares, in a similar connection, with very satisfactory results. See Everett, Parsons, O'Rourke, and Sáenz, "Progress reports on stresses in insulators," Eng. Research Inst. Publ. (December, 1947), and also reference 4.

planned to carry out systematic experimental investigations with symmetrically quenched, unannealed glass cylinders to confirm this point.

III. MEASUREMENT OF n_0

Since it is necessary to know n_0 to calculate the stress distribution σ_ρ, σ_φ, and σ_z from the above integral formulas, a method for determining it will be presented at this point.

The following procedure can be used to measure n_0. One rotates the cylinder about the x axis by an angle α, so that the light strikes it obliquely, and determines the absolute retardations of the rays located in the central section S, as shown in Fig. 2. One also finds the value of the absolute retardations for waves whose planes of vibration coincide with the yz plane and are incident normally on the cylinder, as in Fig. 1(b). This information determines n_0 completely.

To prove this statement, one proceeds in the following way. In section S

$$\left.\begin{aligned} P' &= \sigma_z \cos^2\alpha + \sigma_\rho \sin^2\alpha \\ Q' &= \sigma_\varphi \\ R' &= \sigma_z \sin^2\alpha + \sigma_\rho \cos^2\alpha \end{aligned}\right\}, \tag{13}$$

where P' and Q' are the secondary principal stresses in the plane of the wave front in section S (Fig. 2), and R' is the stress normal to the wave front. Since it can be shown that no rotation of the secondary principal axes occurs along the light path LL' in S, the absolute retardations $R_P^{(\alpha)}$ and $R_Q^{(\alpha)}$ (the superscripts refer to the angle α of oblique incidence) corresponding to plane polarized waves whose planes of vibration are parallel to P' and Q', respectively, are given by

$$\left.\begin{aligned} R_{P'}^{(\alpha)} &= 2\sec\alpha\left\{\int_a^b [C_1(\sigma_z\cos^2\alpha+\sigma_\rho\sin^2\alpha) \right. \\ &\quad \left. + C_2(\sigma_\varphi+\sigma_z\sin^2\alpha+\sigma_\rho\cos^2\alpha)]d\rho + n_0(b-a)\right\} \\ R_{Q'}^{(\alpha)} &= 2\sec\alpha\left\{\int_a^b [C_1\sigma_\varphi + C_2(\sigma_\rho+\sigma_z)]d\rho \right. \\ &\quad \left. + n_0(b-a)\right\} \end{aligned}\right\}, \tag{14}$$

where the factor $\sec\alpha$ arises from the fact that the light paths are longer for oblique than for normal incidence.

The terms in (14) containing σ_φ vanish because of the relation

$$\int_a^b \sigma_\varphi(\rho)d\rho = b\sigma_\rho(b) - a\sigma_\rho(a), \tag{15}$$

which follows from (6). The right-hand side is zero for both solid and hollow cylinders since the boundaries are free from external stresses. From (14) and (15),

$$\left.\begin{aligned} n_0 = \frac{1}{2C(b-a)}\{ & C_2 R_{P'}^{(0)}(\cot^2\alpha - 1) + [-C_2 R_{P'}^{(\alpha)} \\ & + (C_1+C_2)R_{Q'}^{(\alpha)}\sin^2\alpha]\cot\alpha\csc\alpha\} \end{aligned}\right\}, \tag{16}$$

where $R_{P'}^{(0)}$ is the absolute retardation through the section S (Fig. 2) along the P' direction for normal incidence ($\alpha = 0$).

Equations (8), (11), (12), and (16) permit one to determine σ_ρ, σ_φ, and σ_z at all interior and boundary points of symmetrically quenched solid cylinders and concentric hollow cylinders from purely optical data.

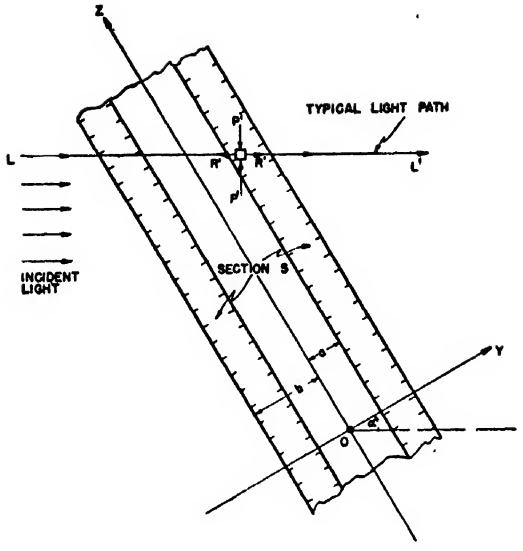


FIG. 2. Oblique incidence on central section S of cylinder (x axis and Q' perpendicular to plane of paper).

IV. CONCLUDING REMARKS

The method presented in the previous sections should be of particular interest for the analysis of residual stresses in glass objects which are so highly quenched that their explosive characteristics would prevent the use of cutting techniques for the measurement of these stresses. It also appears promising in the study of transient thermal stresses in cylinders, particularly for those cases when the temperature differences are so large that thermal and elastic constants of the material vary considerably, and the radiation at the surface does not follow the simple Newton law of cooling, so that the Stefan-Boltzmann law must be employed. Such thermal stresses cannot be determined readily by present photoelastic techniques.

Some Crossover Properties in the Electron Immersion Objective

L. JACOB*
The University, Manchester, England
(Received January 31, 1950)

It was earlier established that the lens field between the cathode and crossover can be represented by a simple function $V=A\sinh kz$, so that the main properties of the paraxial "crossover" (or minimum section of the beam) can be determined as a first approximation in terms of "A" and "k." The distance of the crossover works out at $8\pi/3k\sqrt{3}$; its potential is $A\exp(8\pi/3\sqrt{3})$; and the angle of the beam is simply proportional to "k." The radius of the crossover is then $2/k\exp[-(2\pi/3\sqrt{3})]\cdot\tan\theta$ for electrons of initial velocity "A," where θ is the angle of emission within the paraxial limit. The differential analyzer was also used to obtain the correct solution for the trajectory equation from the field plots for different systems, and shows that the simple theory gives an adequate approximation for design purposes.

I. INTRODUCTION

IT was only with the advent of the study of electron motions in E. S. Fields that their property of crossing over to form a minimum beam section was recognized. The optical counterpart has no practical significance, since focusing of the object is the chief aim, and in a sense it was hardly ever considered. In their observations on the focused spots of television C.R. tubes, Maloff and Epstein[1] concluded that the "object" which was imaged on the screen was a "crossover" produced by the first lens (immersion objective).

This is brought about in the following way. The electrons, which leave the cathode at all angles and with a range of initial velocities, are focused by the immersion lens and produce an image of the cathode. This is usually enlarged and inverted. The "principal rays" however, cross the axis at a point nearer the cathode; the plane perpendicular to the axis at this point contains the "crossover." The envelope of the beam at the crossover gives the minimum cross section of the beam, which is ultimately focused into the small spot on the screen.

The immersion objective consisted of an oxide emitting cathode C, a modulating electrode M, and an accelerator A (Fig. 1). The distance between C and M was varied keeping the position of A fixed, and the field distribution determined in the electrolytic tank. It has been shown[2] that a simple empirical formula $V=Ae^{kz}$ approximately describes the relation between the potential V at any point on the axis and the distance "z" from the cathode, for the region up to and somewhat beyond the modulating electrode, which was explored in the tank. Using this, an approximate theory of crossover properties is presented in what follows, which can be helpful in design. This indicates the importance of the geometrical arrangement of the electrodes.

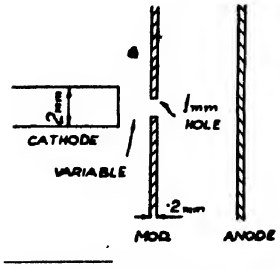


FIG. 1. Immersion objective.

II. THE CATHODE/CROSSOVER DISTANCE

For the purposes of the subsequent analysis, the simple relation $V=Ae^{kz}$ has been used. Here "k" is a geometrical scale factor which varies from geometry to geometry, while "A" is a voltage scale factor, which, for any one geometry and fixed field form, varies correspondingly with the over-all voltage applied to the system. In order to satisfy the boundary conditions at the cathode ($V=0$), the distribution in the region between the cathode and the nearest point of the experimental plot should be represented by $V=A(e^{kz}-e^{-kz})$. The extent of the linear part of the field given by the last expression is then assumed small in comparison with the exponential portion Ae^{kz} which includes the crossover region.

Electrons which are emitted normal to the cathode in the linear field remain normal till they enter the exponential region. Here, they experience the focusing effect. By assuming that the function $V=Ae^{kz}$ applies to this region, the electrons which leave the cathode at $z=0$ must enter the exponential field with a velocity of "A" volts. For paraxial systems this has a value of about 0.5 for a 30-v blackout, and is thus of significance in practice.

We begin with the paraxial form of the trajectory equation:

$$d^2r/dz^2+\tfrac{1}{2}(V'/V)dr/dz+\tfrac{1}{4}(V''/V)r=0 \qquad (1)$$

and, substituting $V=Ae^{kz}$, this yields

$$r=C\exp\{[-k/4+(k\sqrt{3}/4)\cdot i]z\}+D\exp\{[-k/4-(k\sqrt{3}/4)i]z\} \qquad (2)$$

as the complete solution. This gives the height "r" of the ray at any point

$$r=e^{-(k/4)z}\{M\cos(k\sqrt{3}/4)z+N\sin(k\sqrt{3}/4)z\}. \qquad (3)$$

To evaluate the constants M and N, we must put in the

* Now at University of Liverpool, England.
[1] I. G. Maloff and D. W. Epstein, *Electron-Optics in Television* (McGraw-Hill Book Company, Inc., New York, 1938).
[2] P. A. Einstein and L. Jacob, Phil. Mag. 39, 20 (1948).

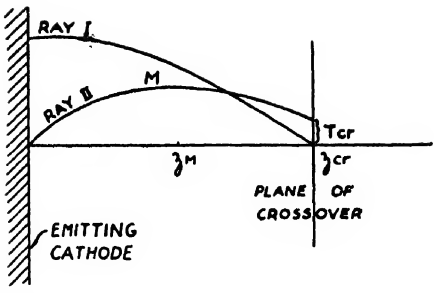


FIG. 2. Rays defining crossover and its plane.

boundary conditions. An electron starting at the top of the emitting area and moving parallel to the axis is deviated by the field so as to cut the axis at the crossover (Fig. 2); this gives one focal plane of the immersion objective. Now put $r=a$, the radius of the emitting area when $z=0$; hence $M=a$. Also $(dr/dz)_{z=0}=0$; hence

$$N=a/\sqrt{3}; \tag{4}$$

$$\therefore \quad r=ae^{-(k/4)z}\{\cos(k\sqrt{3}/4)z+(1/\sqrt{3})\sin(k\sqrt{3}/4)\cdot z\}. \tag{5}$$

Now, let this trajectory meet the axis in the crossover plane, at a distance z_{cr} from the cathode. Here, $r=0$.

$$\therefore \quad \cos(k\sqrt{3}/4)\cdot z_{cr}+(1/\sqrt{3})\sin(k\sqrt{3}/4)\cdot z_{cr}=0; \tag{6}$$

$$\therefore \quad z_{cr}=8\pi/3k\sqrt{3}. \tag{7}$$

Thus, the crossover distance from the cathode varies inversely as "k," the geometrical scale factor.

Actually, the value of the crossover distance determined from (7) will be a limiting value, since it corresponds to electrons with initial speeds of about 0.5 v. The most probable crossover corresponding with the most probable velocity of about 0.1 volt will lie somewhat closer to the cathode, but the difference is only some five percent and may be overlooked for this type of system. This "chromatic aberration" effect may however be important in systems where there is a wide spread in z_{cr} or of initial velocity; its effect will be most marked when crossover diameters are to be derived.

III. THE BEAM ANGLE AND POTENTIAL AT THE CROSSOVER

The immersion objective has the form of a "thick lens," i.e., consists of a series of contiguous refracting layers, so that an electron ray is continuously refracted all through the field. The direction in which it emerges and which defines the angle of the beam, as measured in a field free space outside the lens, is usually very different from the angle φ at which it crosses the axis at the crossover. The angle φ is derived by differentiating (5) and putting

$$z=z_{cr}=8\pi/3k\sqrt{3};$$

$$\therefore \quad \tan\varphi=-(ak/2)\exp(-2\pi/3\sqrt{3}). \tag{8}$$

This gives the angle at the crossover, in terms of the height at which the ray left the cathode normally, and the "k" of the field. For a constant emitting area of radius "a," the angle at the crossover is directly proportional to the factor "k" which defines the geometry of the system. It is, of course, also proportional to the diameter of the coated area.

For a given geometry, the beam angle is thus constant, independent of the total voltage on the system, because the factor "A" does not determine its value. This is a well known experimental law which is used in the Electron-optical voltmeter.[3]

The potential on the axis at the crossover is simply derived by substituting the value of z_{cr} from (7) in the equation $V=Ae^{kz}$ for the axial distribution. This yields a value $A\exp(8\pi/3\sqrt{3})$, which is thus proportional to A; hence, A is some constant fraction of the crossover voltage. Under the approximation used, it is also the initial velocity with which the electrons enter the exponential part of the field, as stated previously. If the total voltage on the system is increased "n" times, that at the crossover also increases "n" times, and the ratio of crossover potential to final potential is not altered. Hence, no change would be expected in the magnification of an image of the crossover, when the total voltage on the tube was varied in any way, apart from the effect on the space charge. Use can be made of this fact by finding

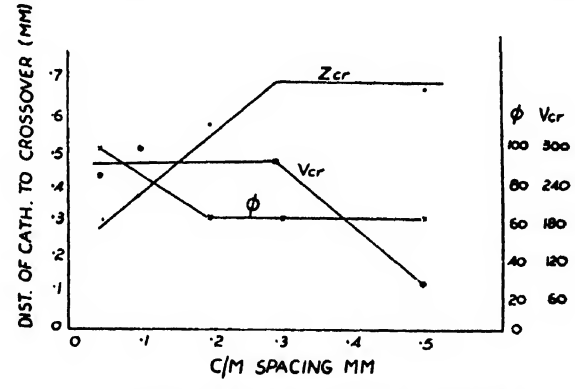


FIG. 3. Effect of spacing on crossover parameters.

the point at which the relationship no longer holds; the control of crossover size by space charge then steps in, and the magnitude involved can be assessed.

The variation of these crossover properties with C/M spacing is shown in Fig. 3. These were derived from the field plots for the different systems. They include crossover distance z_{cr}, angle at the crossover φ and its potential V_{cr} calculated from the above formulae using the constants determined from each geometry. The data has been converted to correspond with the case of a vacuum tube system run under the condition of constant anode voltage for each system. It will be seen that the potential at the crossover remains nearly constant except for the largest spacing. Also, the beam angle decreases and appears to reach a limiting value of 60° at

[3] L. Jacob, J. Inst. Elec. Eng. 91, 512 (1944).

this spacing. In an actual sealed off tube, the largest angle at the crossover is about 100°. In practice, using a fluorescent disk to measure the angle, the measured angles are only about 20°, i.e., one-fifth the calculated values. This arises from the fact that the angle is measured in a field-free space at the final anode voltage, while the angle at the crossover is situated in a space at a very much lower potential. Care should be taken to allow for this difference, when using measured values in any calculations which have to be made.

IV. THE EFFECT OF MODULATION ON CROSSOVER PROPERTIES

The field plotting technique can be readily applied to follow through the effect of modulating the beam, by applying negative potentials, V_m, to the modulator,

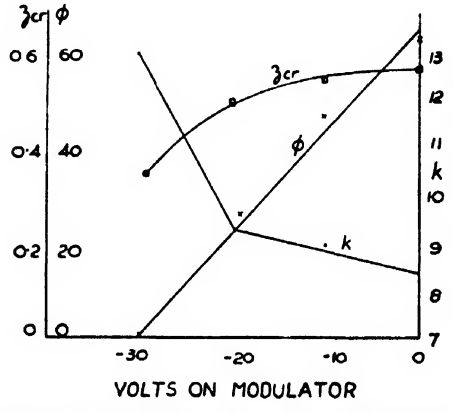


FIG. 4. Effect of modulation on crossover parameters.

with respect to the cathode. The family of curves obtained for negative voltages up to the blackout point follow the same exponential law as for $M=0$, but the value of "A" decreases while that of "k" increases in each case. A separate experiment in which the emitting area of the cathode was covered with a square gauze network showed clearly that one effect of the modulation was to reduce the radius of the emitting area of the cathode in proportion to the modulation swing. For example, if the system required -30 volts to black out the beam, then if -10 volts was applied to the modulating electrode, the emitting radius was restricted to 20/30 of its size when the modulator was at zero or cathode potential. Hence "a" in (8) decreases in proportion to the modulation, while "k" usually increases somewhat (Fig. 4).

The product "ak" which determines the angle is predominantly affected by the large variation in the emitting radius, and this is the controlling factor, so that the calculated angle φ varies almost linearly with V_m.

There is a further effect as regards the position of the crossover, z_{cr}, when the beam is modulated; it tends to move toward the cathode. This shift would be expected on purely qualitative grounds, since it is known that the equipotentials become more and more curved as the inter-electrode space becomes more negative. This, in turn produces larger deviations of the rays, i.e., shorter focal lengths; the crossover point will thus tend to move towards the cathode. With this type of system, the crossover remains on the side of the modulator remote from the cathode throughout the modulation range.

V. THE SIZE OF THE CROSSOVER

The intersection of the path of an electron which leaves the cathode at normal incidence and with zero velocity (at some height above its center) with the axis of the system defines the plane of the crossover for this velocity. All such electrons emerging at different heights will meet at this point under truly paraxial conditions. The upper limit to the size of the crossover will be defined by the group of electrons leaving the center of the cathode at an angle of 90°, and with the greatest velocity. The points at which these strike the plane of the crossover sets the limit to its boundary. The actual physical number involved here will be small since the emitter can be assumed to obey the Lambert cosine law. Further, for each direction of emission, the Maxwell distribution will determine the fraction of the total number with any assigned velocity; for example about 70 percent of the electrons are to be found with all velocities up to the most probable velocity of about 0.1 volt. It is therefore sufficient to examine the crossover sizes obtained for electron speeds of 0.1 volt as a working limit, since the area enclosed will embrace some 70 percent of all the electrons emitted by the source. Unfortunately, it was not possible to adhere to this value for the simple theory, since the initial electron velocity is fixed by the value of A, and this is about 0.5 v when the blackout voltage is 30 v. Even with this limitation, the theory is worth stating since it provides data for an upper limit to the size of the crossover. The size of the final image can be calculated from this, once the potential in the crossover plane is known.

In order to trace the rays which leave the center of the cathode at angles within the paraxial range, we use the original trajectory equation, but the constants will have different values because of the new boundary conditions. The initial linear field is again neglected here, but a rough calculation shows that by the time an electron has passed through it, its early parabolic path has now a slope within the paraxial range of the exponential focusing field, and the crossover height will not be affected by the height at which the ray starts. Since

$$r=e^{-(k/4)z}\{P\cos(k\sqrt{3}/4)z+Q\sin(k\sqrt{3}/4)z\}$$

from (3) the value of P is zero when $r=0$; for $z=0$.

$$\therefore\quad r=e^{-(k/4)z}\cdot Q\sin(k\sqrt{3}/4)z. \tag{9}$$

When $z=z_{cr}$, then $r=$radius of the paraxial crossover $=r_{cr}=\exp[-(k/4)z_{cr}]\cdot Q\cdot\sin(k\sqrt{3}/4)z_{cr}$. Now put $z_{cr}=8\pi/3k\sqrt{3}$

$$\therefore\quad r_{cr}=\exp[-(2\pi/3\sqrt{3})]\cdot Q\sin(2\pi/3). \tag{10}$$

This ray will pass through the periphery of the crossover. Before doing so, it may proceed to some point corresponding to z_M along the axis at which its height is maximum, while being focused toward its image on the axis. At this maximum height

$$(dr/dz)_{z=z_M}=0;$$

hence, from (9),

$$z_M=(4\pi/3k\sqrt{3})=z_{cr}/2. \tag{11}$$

This ray thus reaches its maximum height at a distance from the cathode equal to half the crossover distance. The height of the ray at this point is $\exp(-\pi/3\sqrt{3})\cdot Q\cdot\sqrt{3}/2$ and will obviously be governed, apart from its initial velocity "A," by the initial angle θ (assumed within the paraxial range) at which the electron leaves the cathode center, i.e., $(dr/dz)_{z=0}=\tan\theta$. Hence from (9),

$$Q=4\tan\theta/k\sqrt{3}; \tag{12}$$

$$\therefore\quad r=\exp[-(k/4)z]\cdot 4\tan\theta/k\sqrt{3}\cdot\sin(k\sqrt{3}/4)\cdot z. \tag{13}$$

This gives the height of the ray at any point in terms of the angle of ejection, and the geometry of the field

$$\therefore\quad r_{cr}=(2/k)\cdot\exp[-(2\pi/3\sqrt{3})]\cdot\tan\theta. \tag{14}$$

Hence, the paraxial crossover size is determined for electrons of initial velocity "A" by the ratio of $\tan\theta$ to "k." It might be supposed that there was no limit to the crossover size if θ took on the value of 90°. The theory would of course not hold for this case. Actually, the work with the differential analyzer described in the next section shows how quickly such rays are brought in by the field; the modulator hole does not appear to act in any way as a limiting stop. An examination of (14) indicates that the crossover size can be reduced only by a variation in geometry. In the standard C.R. tubes, "k" has already a value of about 10, and it is unlikely that it can be made much smaller by increasing the geometrical factor alone beyond this limit.

The aim in design should be to increase "k" to as large a value as possible, while at the same time insuring that the crossover is situated at a point in the field which has the lowest potential.

This should yield the minimum crossover image at the screen in theory; in practice, lens aberrations and space charge appear to be the controlling factors.

The maximum height of the trajectory was shown to be $\exp(-\pi/3\sqrt{3})\cdot Q\cdot\sqrt{3}/2$ and this occurs at a point approximately halfway to the crossover. All the rays leaving the center of the cathode will reach maximum height in the vicinity of this point; the peak height is found on substitution to be about 1.8 times the height of the crossover, independent of θ.

VI. TRAJECTORIES WITH THE DIFFERENTIAL ANALYZER

The crossover properties deduced from the simple theory can be helpful in design of systems, and the approximate form of the trajectory can also be traced out point by point for the paraxial case from (13).

In order to deal with electrons leaving the center of the cathode at angles *outside* the paraxial range, i.e., up to 90°, the differential analyzer was employed to solve the trajectory equation.

The general equation has been stated (1) as follows:

$$d^2r/dz^2+\{[1+(dr/dz)^2]/2(V+A)\}\cdot\partial V/\partial z\cdot dr/dz -\{[1+(dr/dz)^2]/2(V+A)\}\cdot\partial V/\partial r=0, \tag{15}$$

where V is the potential at any point on the axis and "A" is the initial electron velocity.

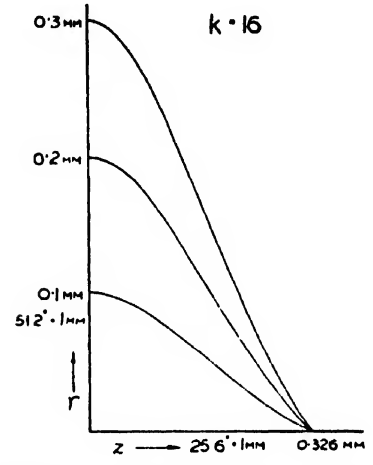


FIG. 5. Trajectories with differential analyzer.

This equation is not easily manageable unless paraxial conditions are imposed, e.g., $(dr/dz)\ll 0.1$, in which case it takes the form

$$d^2r/dz^2+\{V'/2(V+A)\}dr/dz +\{V''/4(V+A)\}\cdot r=0. \tag{16}$$

The differential analyzer solutions were, however, obtained from (16) which was taken in the simpler form of the two equations of motion, and before the time parameter "t" had been eliminated, i.e.,

$$m(d^2z/dt^2)=e(\partial V/dz), \tag{17}$$

$$m(d^2r/dt^2)=-er(\partial^2V/\partial z^2)/2. \tag{18}$$

If the electrons leave the cathode with initial velocity "A," then the results of the analyzer traces show that they follow approximately the same trajectory under the potential field $V=(Ae^{kz}-A)$, as those with zero velocity under a field $V=A(e^{kz}-e^{-kz})$. The value of "A" for cases of 30-v blackout is, as stated earlier, about 0.5 volt. This gives the diameter of the crossover calculated from (14) a practical significance. The analyzer traces were taken assuming the field to have the sinh-form all the way from the crossover to the cathode surface. The periphery of the crossover is then defined by the electrons which leave the center of the cathode at 90° and with the most probable velocity of 0.1 v.

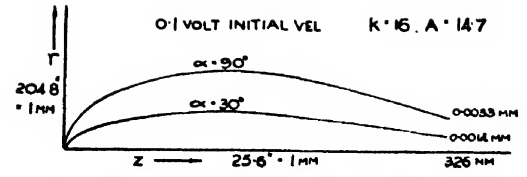


FIG. 6. Trajectories defining radius of crossover.

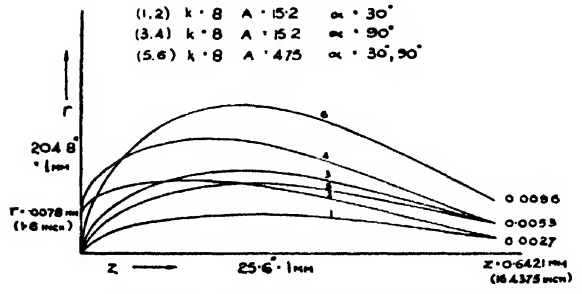


FIG. 7. Trajectories for 30° and 90° emission.

The electron distribution at the crossover assuming ideal focusing systems has already been treated by Langmuir[4] and will not be further discussed here.

In addition to the 90° trajectory, a further one of 30° was taken as a matter of interest, using the most probable velocity of 0.1 volt for the electrons leaving the center of the cathode. In order to find the crossover plane, it was necessary to trace electrons leaving the cathode normally at different heights above the axis. These are shown in Fig. 5 for a system with a value of $k=16$ and 5000 volts on the anode, and for an initial velocity of 0.1 volt. In order to keep within the paraxial region, the largest height of incidence used was 0.3 mm (emitting radius 0.5 mm). For the same system, the limiting value of the crossover radius is seen from Fig. 6 to be 0.0033 mm.

The size of the crossover will, for a given initial velocity, depend on the voltage applied to the system. If this is varied, the size is readily determined from the machine constants. For a 30-v blackout with this system and a 0.1 volt emission its value is 0.033 mm.

Figure 7 shows a series of trajectories for a system having $k=8$ and for various anode voltages, for angles

[4] D. B. Langmuir, Proc. I. R. E. 25, 977 (1937).

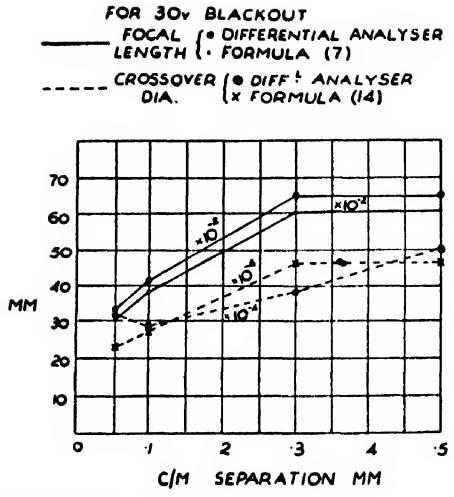


FIG. 8. Comparison between differential analyzer and simple theory.

of 30° and 90°. It will be observed from the pairs (1) and (2), and also (3) and (4), which relate to rays leaving the cathode surface with the same direction and velocity, but at different heights, that they are focused to the same point in the crossover plane, which is to be expected under these conditions. It should also be noticed that the trajectories from the center of the cathode reach a maximum height at a point roughly halfway to the crossover, and this is usually about twice the crossover height, as shown earlier.

The focal length of the lens formed by the immersion objective system is seen from Fig. 8 to increase with the C/M separation. The value calculated by formula (7) is shown to be a good approximation when compared with the traced value on the analyzer. The same figure shows how the crossover diameter varies as the cathode is moved away from the modulator keeping the blackout voltage constant at the value of 30 v. The approximation obtained from (14) is shown to be quite good, even though no rays are included which are inclined at an angle greater than 17°.

In conclusion, I wish to thank the Director of the National Physical Laboratory for his kindness in placing the facilities of the differential analyzer (housed in this University) at my disposal, and Messrs. Michel and Butler for helpful discussions during its operation.

Single Crystal Copper Surfaces*

T. N. Rhodin, Jr.
Institute for the Study of Metals, University of Chicago, Chicago, Illinois
(Received February 20, 1950)

Some surface structure characteristics of a metal single crystal are considered and the conclusions are applied to the preparation of single crystal plates of copper. A successful preparation is described as well as suitable methods for evaluating the effectiveness of the preparation. Using physical microadsorption isotherms of nitrogen at 78.1°K, electron diffraction, and x-ray diffraction methods, the conclusion is tentatively asserted that the surfaces are essentially planar on a molecular scale, and that the surface atoms are arranged in an approximately undistorted lattice.

I. INTRODUCTION

It has become a question of great interest to ascertain the relationship between the physical and chemical properties of a metal interface and that particular crystallographic arrangement characteristic of the atoms in the interface area between the metal and its environment. The environment could be a fluid or solid such as another grain of the same metal. Although these remarks apply equally well to any such interface in general, let us for convenience limit consideration to a single crystal-gas interface. The surface of the metal crystal may, in a region where the macroscopic surface has a given orientation, approximate the ideal crystal plane of the corresponding orientation—that is to say, a surface formed from an infinite crystal lattice by removing all atoms whose centers lie on one side of the plane. This would give steps and risers of atomic dimensions only.[1] The surface may, on the other hand, have a hill and valley structure on a scale larger than atomic dimensions,[2,3] so that the planar arrangement of the microscopic surface may well differ from that of the macroscopic surface as illustrated in Fig. 1. Thermodynamic considerations prescribe that the surface structure with the lowest free energy will be the true one providing that the surface be in equilibrium with its vapor for a sufficiently long time. However, as is so often the case for the solid state, since the structure actually obtained after a finite time will depend on the rate of approach of the system to an equilibrium state, the conditions experimentally obtained at room temperature may differ considerably. Hence, although there is much evidence[4] for the unlikelihood of hill and valley structures and for the validity of the hypothesis that surfaces are atomically smooth if the temperature is sufficiently high, such observations do not necessarily apply to metal surfaces prepared at low temperatures. Yet the definition of surface structure is a critical prerequisite to surface studies and it is exceedingly difficult to do so, when room temperature surface polishing techniques tend to distort mechanically the structure of the surface layers.[5] On the other hand, if the surface distortion were minimized the surface orientation of sufficiently planar surfaces would be that of the matrix atoms. The preparation of copper surfaces of this type suitable for surface chemistry studies using a sensitive quartz vacuum microbalance was systematically investigated without resorting to high temperature equilibrium conditions. Weight increments caused by surface effects such as adsorption or oxidation are often small and, even for an exceedingly sensitive instrument, it is desirable for the samples to possess a high surface to volume ratio; in other words, the optimum geometry of the sample is that of a very thin plate. It is difficult enough to satisfy the conditions of planarity and undistorted surface structure under the most favorable conditions but special considerations are required when this third condition is added. Recently a systematic study of these considerations, applicable to thin single crystal plates of copper, was completed. A concise report of the results seems desirable.

II. PROCEDURE AND RESULTS

The original source of the crystal plates was a cylindrical single crystal of copper (6 in.×1¼ in.) grown by

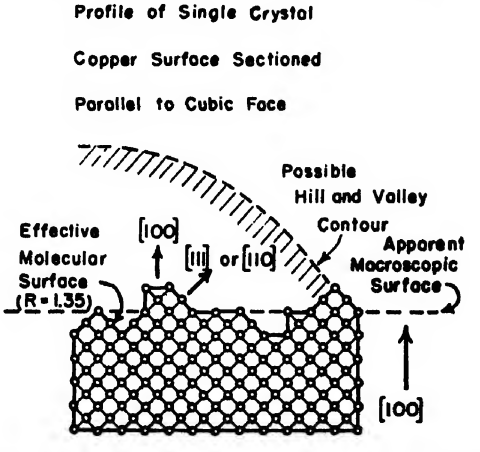


Fig. 1. Profile of single crystal copper surface sectioned parallel to cubic face (schematic). (Arrows indicate crystallographic directions normal to crystal face.)

* The research was supported in part by Army Air Forces Contract AF-33(038)-6534.
[1] S. T. Martin, Phys. Rev. 56, 947 (1939).
[2] I. Langmuir, Phys. Rev. 22, 357 (1923).
[3] L. Tonks, Phys. Rev. 38, 1030 (1931).
[4] C. Herring and M. H. Nichols, Rev. Mod. Phys. 21, 258 (1949).
[5] C. F. Elam, Trans. Faraday Soc. 32, 1604 (1936); S. Dobinski, Nature 138, 685 (1936).

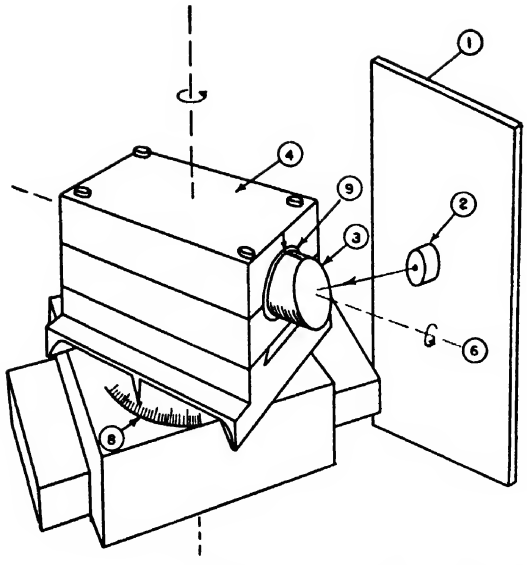


FIG. 2. Drawing of jig holder for orientation determination of cylindrical single crystals. (1) Film cassette; (2) x-ray pinhole; (3) crystal; (4) jig; (5) vertical axis rotation; (6) horizontal axis rotation; (7) cutting wheel; (8) graduated circle; (9) fiber spacer.

the standard Bridgman method. The orientation and symmetry of the crystal was determined by the Laue method, using back reflection x-rays. Testing the crystal at various spots by this method indicated an over-all variation of lattice orientation of two degrees or less. The copper was 99.995 percent pure with traces of iron, silicon, and aluminum. Sections ($\frac{3}{4}$ in$\times\frac{1}{2}$ in.) were cut from the crystal very slowly with an alundum cutting wheel at a proper angle so as to expose the macroscopic face along a predetermined low index plane. The proper transfer of the crystal from the Laue camera set-up to the cutting wheel was insured by the use of an (some of the techniques roughly resemble those used by A. T. Gwathmey and A. F. Benton, J. Phys. Chem. 44, 35 (1940)) interchangeable jig arrangement schematically indicated in Figs. 2 and 3, respectively. The crystal was so arranged in the jig to permit one to orient the pole of the desired plane by simple rotations about two axes, one parallel to the cylinder axis and one normal to it, as indicated by items 5 and 6 in Fig. 2. Mounting the section in Lucite with one face exposed permitted one to polish it mechanically by the usual methods, finishing up with Buehler powder No. 3 on the polishing wheel. Upon reversal in the mount the second face was likewise polished until the thickness of the sample had been reduced to 0.040 in. This preliminary mechanical polish was necessary to insure uniform removal of material by the relatively prolonged electropolishing that followed. The plate was thereupon floated off the mount in benzene and the thickness reduced 75 percent to 0.010 in. by a carefully tedious process of electropolishing so as to insure removal of the distorted layers. The technique in this stage was a standard one except that the rate of removal of material was kept very small. A solution of orthophosphoric acid and ethyl alcohol was used in the usual proportions with an etching current of less than 0.01 ampere at about a potential of 2.0 volts for several periods of 20 minutes each until sufficient material had been removed. The long treatments tended to develop a wavy surface, but on such a large scale that it did not contribute to the molecular roughness. The final appearance was mirror bright and planar. No pitting nor selective attack of the surface was observed by microscopic observation at 1000 diameters.

The effectiveness of the preparation was tested in two ways. The planarity of the final surfaces on a molecular scale was determined from microadsorption isotherms of nitrogen at 78.1°K using a very sensitive microbalance technique described in detail elsewhere.[6] In brief, the amount of physically adsorbed gas as a function of pressure was measured and the weight of adsorbed gas corresponding to a monolayer coverage of the surface by a close packed layer of adsorbed gas molecules was calculated from the adsorption isotherms by the Brunauer-Emmett-Teller equation[7] and checked by the Harkins-Jura method.[8] Using the proper value for the cross-sectional area of an adsorbed nitrogen molecule a simple calculation yields the value of the effective surface area available for nitrogen molecule adsorption, that is to say, the molecular surface area. The ratio of this surface area to that calculated from the macroscopic dimensions of the sample gives a precise value for the roughness of the surface. For example, if (R) is the roughness factor, the percent roughness of the surface on

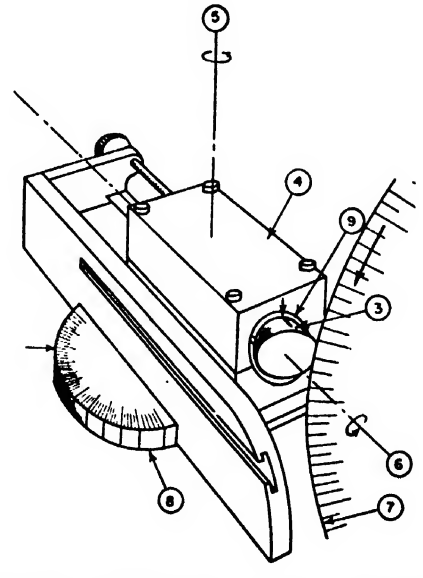


FIG. 3. Drawing of jig holder for sectioning of cylindrical single crystals (1) Film cassette; (2) x-ray pinhole; (3) crystal; (4) jig; (5) vertical axis rotation; (6) horizontal axis rotation; (7) cutting wheel; (8) graduated circle; (9) fiber spacer.

[6] T. N. Rhodin, J. Am. Chem. Soc. (to be published).
[7] S. Brunauer, *The Adsorption of Gases and Vapors*, Vol. I, *Physical Adsorption* (Princeton University Press, Princeton, New Jersey, 1942), p. 149 ff.
[8] W. Harkins, and G. Jura, J. Am. Chem. Soc. 66, 1366 (1944).

a molecular scale may be expressed as $100(R\text{-}1)$. Since it is reasonable to assume for a metal surface that irregularities will tend to be more or less spread out over the whole surface, surfaces with roughness factors near unity will be essentially planar. Such was the case for the surfaces prepared as described, where most of the R values ranged from 1.2 to 1.5.

The perfection of the surface structure was examined directly by both glancing angle electron diffraction and transmission and back-reflection of x-rays. No definable diffraction pattern was obtainable from the electron diffraction examinations, although the surface had been carefully cleaned by distilled water and alcohol rinses followed by a 400 C. hydrogen reduction and outgassing pre-treatment. Back-reflection x-ray patterns indicated a maximum lattice distortion of two degrees averaged over approximately the outer 5000 layers of copper atoms. In order to define better the structure of the surface layers, transmission x-ray examinations were made of a small section in the center of a crystal plate which was systematically reduced in thickness by successive electropolishing treatments. The ultimate thickness examined appeared to be less than 10^{-4} cm with no indication that the observed distortion was not more or less characteristic of the whole thickness of the plate.

III. DISCUSSION

If one assumes that the surfaces used for electron diffraction examination were adequately clean, the inability to observe coherent reflection patterns may just indicate the absence of suitable projections on the surface for the electrons to penetrate without being totally scattered. This might be conceivably true for surfaces approximately planar on a molecular scale. In addition, x-ray reflection and transmission patterns, indicating a maximum variation in lattice perfection of two degrees, may be considered to indicate a relatively undistorted surface layer providing one accepts the admittedly inconclusive indication that the observed small distortion is not intensively concentrated in the surface layers. It cannot be denied that the results do not preclude the possibility that a highly distorted outermost layer, too thin to contribute to the x-ray pattern, may exist. In addition, well-defined reflection patterns have been obtained for glancing angle electron reflection from planar surfaces of mica and from thin films of dielectrics formed by condensation on to mica substrates.[9] However, the existence of this effect for unetched single crystal metal surfaces has yet to be established and the present inability to observe it remains an open question.

In the final analysis the fact that atoms are arranged in a particular crystallographic plane on the surface may be best indicated by some physical chemical effect on a molecular scale, the magnitude of which can be measured and the extent of which is sensitively dependent on the particular crystallographic face involved. Two such effects, namely, low temperature oxidation, and low temperature physical adsorption of gases have been investigated in this laboratory. In both cases the magnitude of the energy terms involved, e.g., the activation energy for oxidation and the heat of physical adsorption have been observed to depend on the crystallographic direction of the macroscopic surface as if most of the surface atoms were arranged in a characteristic well-ordered surface lattice.[10] Since both phenomena occur by a molecular mechanism, there is reason to believe that there exists a precise and reproducible dependence on the surface structure. Although a detailed discussion of these effects is out of place here, the fact that they have been observed is cited as evidence for the assumption that the surface structures of the samples prepared, as previously described, correspond to a considerable extent to that of a fairly well-ordered surface lattice.

IV. CONCLUSIONS

It is believed that thin relatively undistorted single crystal plates of copper can be prepared from large single crystals. A suitable process of mounting and electropolishing is described, as well as a sample holder to facilitate orientation determination and sectioning. There is evidence to believe that the surfaces are essentially planar and that the surface atoms exist in an orientation corresponding to a considerable extent to that characteristic of parallel crystallographic planes in the matrix of the crystal.

V. ACKNOWLEDGMENTS

The author is grateful to K. Ikeuye, J. Hess, and L. Schulz for discussions and assistance. The painstaking technique required for the preparation of the samples was furnished by the patient effort of J. Cerny and is also gratefully acknowledged.

[9] L. G. Schulz, J. Chem. Phys. 18, 996 (1950).
[10] T. N. Rhodin, J. Am. Chem. Soc. (to be published).

Theory of the Parallel Plane Diode

A. H. Taub and Nelson Wax
University of Illinois, Urbana, Illinois
(Received March 10, 1950)

Transient and steady state solutions of the parallel plane diode are obtained. The limitations on a single valued velocity theory are investigated. These general results are applied to the space charge limited diode with the a.c. components small compared to the d.c. Power consumption is discussed and it is shown that hitherto neglected terms are of importance.

I. INTRODUCTION

THE parallel plane diode has been the subject of a large number of investigations, both theoretical and experimental.[1–6] A good deal of attention has been centered on problems concerning the behavior of the plane diode when both space charge and time-varying fields are present. These latter studies usually[4–6] have employed perturbation techniques: the time-varying quantities were assumed to be first-order corrections on the static quantities.

It is the purpose of this paper to discuss the theory of the plane diode when a time dependent current density flows between the cathode and plate. The following assumptions are made throughout: (1) The diode consists of an infinite plane cathode parallel to an infinite plane anode (edge effects are ignored). (2) The magnetic effects of the moving electrons are neglected.

Under these assumptions the fundamental equations are reduced to dimensionless form in II. The transient behavior of the diode is treated in III, and it is shown that the differential equations of the system may be solved by quadratures if the initial position, x^0, and velocity, v^0, of every electron is known at the time, $t=0$, when the current starts to flow. The steady state behavior is discussed in IV; the equations may be integrated directly again when ξ, the time of emission of an electron from the cathode, and t, the time, are chosen as independent variables. The necessary and sufficient condition that the electron trajectories shall not intersect is obtained as an immediate consequence of this formulation. The electron transit time and the plate voltage are given as functions of ξ and t, in V and VI, respectively. The treatment in these, and in subsequent sections, is restricted to a small signal analysis of the space charge limited diode. Fourier expansions for all the quantities which were discussed in earlier sections are given in VII, and these expansions are used in VIII to obtain an expression for the average power consumption of the diode per unit area. It is shown that the notion of complex impedance as used by previous workers is insufficient alone to account for all the power consumed by the diode.

All dimensional equations are in m.k.s. units.

II. THE FUNDAMENTAL EQUATIONS

The plane diode is described by the following three differential equations. The equation of motion of an electron in the diode is

$$(d^2x'/dt'^2)=(e/m)E', \tag{2.1}$$

where x' is the distance of the electron from the cathode, t' the time, E' the electric field intensity, and e/m the charge to mass ratio of an electron.

The equation of conservation of charge is

$$\partial(\rho'v')/\partial x'+\partial\rho'/\partial t'=0 \tag{2.2}$$

with ρ' the charge density and v' the velocity of the electrons associated with ρ'. $v'=dx'/dt'$.

Poisson's equation takes the form

$$\partial E'/\partial x'=\rho'/\epsilon_0. \tag{2.3}$$

ϵ_0 is the dielectric constant of free space $=1/36\pi\times10^{-9}$ farad per meter.

The total current density, I, is defined by

$$I=\rho'v'+\epsilon_0(\partial E'/\partial t'). \tag{2.4}$$

The first term in the right-hand member of Eq. (2.4) is the conduction current density and the second is the displacement current density. It is of particular note that I is a function of t' alone, and not of x'. This result may be obtained by substituting ρ', from Eq. (2.3), into Eq. (2.2) and using the definition for I (see Eq. 3.14).

If I is a periodic function of t', with period τ, then one can introduce the dimensionless set of variables

$$x=x'/a, \tag{2.5}$$

and

$$t=t'/\tau, \tag{2.6}$$

[1] W. E. Benham, Phil. Mag. 5, 641 (1928); 11, 457 (1931).
[2] J. Müller, Hochfrequenztech. u. Elektroakustik 41, 156 (1933); 43, 195 (1934).
[3] F. B. Llewellyn, *Electron Inertia Effects* (Cambridge University Press. London, England, 1943), contains a bibliography.
[4] H. W. König, Hochfrequenztech. u. Elektroakustik 62, 76 (1943).
[5] F. Gray of the Bell Telephone Laboratories Inc. has obtained the general solution of the diode equations by a different method in an unpublished memorandum dated 11/11/41. This material was presented as an invited paper "The general solution for an electron stream in a parallel plane diode" at the 266th meeting of the Am. Phys. Soc. June 15–16, 1945. Gray did not interpret his solutions in the fashion presented here. König (reference 4) has used an approach which is similar to, but more restricted than, Gray's in discussing the system of differential equations (2.11)–(2.13).
[6] H. W. König, *Laufzeittheorie der Elektronenrohren* (Verlag. Julius Springer, Vienna, 1948), Vol. 1.

where a is the distance between the cathode and plate of the diode.

The total current density may now be written as

$$I = I_0[1+\mu f(t)] = I_0 h(t). \tag{2.7}$$

I_0 and μ are constants, and $f(t)$ is an arbitrary periodic function of period one in t (τ in t'), and of maximum absolute value one. I_0 is thus the direct current density, and μI_0 is the maximum amplitude of the time dependent current density.

The dimensionless quantities v, E, and ρ are defined by

$$v = (\tau/a)v', \tag{2.8}$$

$$E = (\epsilon_0/\tau I_0)E' \tag{2.9}$$

and

$$\rho = (a/\tau I_0)\rho', \tag{2.10}$$

respectively. Equations (2.1) to (2.3) may be written in dimensionless form as

$$d^2x/dt^2 = (6/\kappa^3)E, \tag{2.11}$$

$$\partial(\rho v)/\partial x + \partial\rho/\partial t = 0, \tag{2.12}$$

and

$$\partial E/\partial x = \rho, \tag{2.13}$$

respectively. $\kappa^3 = (6am\epsilon_0/eI_0\tau^3)$ is a dimensionless constant.

III. THE TRANSIENT CASE

Let the position x^0 and the velocity v^0 of every electron be known, in the region between cathode and plate, at the time $t=0$, the instant at which the known current I starts to flow. At time $t=t_1$ the electron which emerged from the cathode at $t=0$ will be collected at the plate. It is the object of this section to study the behavior of those electrons which were in the anode-cathode region, at $t=0$, in the time interval $0 \leq t \leq t_1$.

If x^0 is known, then ρ^0, the density distribution of the electrons at $t=0$, is known. Note that $v^0 = v^0(x^0)$ and $\rho^0 = \rho^0(x^0)$. The subsequent position of any electron may be represented by

$$x = x(x^0, t), \tag{3.1}$$

where x^0 is its position at $t=0$.

Equation (3.1) together with the equation

$$t = t \tag{3.2}$$

may be considered as a transformation of coordinates from the variables x, t to the variables x^0, t. The velocity of every electron may be considered to be a function of both x^0 and t. Indeed Eq. (3.1) is the solution of

$$dx/dt = v(x^0, t). \tag{3.3}$$

Furthermore, if one fixes one's attention on any individual electron, then one has, using Eqs. (3.1) and (3.3), that

$$dx/dt = \partial x/\partial t = v(x^0, t), \tag{3.4}$$

since x^0 is fixed for each electron.

The matrix, M, of the Jacobian of the transformation defined by Eqs. (3.1) and (3.2) is

$$M = \begin{Vmatrix} \partial x/\partial x^0 & \partial x/\partial t \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} \partial x/\partial x^0 & v \\ 0 & 1 \end{Vmatrix}. \tag{3.5}$$

The matrix, M', of the Jacobian of the inverse transformation (i.e., the transformation which gives x^0 and t in terms of x and t) is

$$M' = \begin{Vmatrix} \partial x^0/\partial x & \partial x^0/\partial t \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} \dfrac{1}{\partial x/\partial x^0} & -\dfrac{v}{\partial x/\partial x^0} \\ 0 & 1 \end{Vmatrix}. \tag{3.6}$$

Thus, for any function $F(x, t)$ one has

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial x^0}\frac{\partial x^0}{\partial x} = \frac{\partial F}{\partial x^0}\frac{1}{\partial x/\partial x^0} \tag{3.7}$$

and

$$\frac{\partial F}{\partial t} = -\frac{v}{\partial x/\partial x^0}\frac{\partial F}{\partial x^0} + \frac{dF}{dt}. \tag{3.8}$$

The notation dF/dt used in Eq. (3.8) is to denote the derivative of F with respect to t keeping x^0 constant, i.e., following a particular electron along its path. Furthermore,

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + v\frac{\partial F}{\partial x}, \tag{3.9}$$

when the variables x and t are used.

The equation of the conservation of charge is equivalent to the statement,

$$\frac{\partial x}{\partial x^0} = \frac{\rho^0(x^0)}{\rho}, \tag{3.10}$$

if x is a unique function of x^0 and conversely that x^0 is a unique function of x. One assumes, in writing Eq. (3.10), that M, and therefore M', are non-singular. This implies that electron trajectories cannot cross or touch each other.

The Poisson equation now takes the form

$$\partial E/\partial x^0 = \rho^0(x^0), \tag{3.11}$$

and the equation of motion remains unchanged

$$d^2x/dt^2 = (6/\kappa^3)E. \tag{3.12}$$

The general solution of Eq. (3.11) is

$$E = \int_0^{x^0} \rho^0(\eta)d\eta + g(t) + E_0, \tag{3.13}$$

where E_0 is the electric field intensity at $x^0=0$ and $g(t)$ is an arbitrary function of time. The particular solution

for the diode may be obtained by noting that

$$\frac{dE}{dt}=\frac{\partial E}{\partial x}v+\frac{\partial E}{\partial t}=\rho v+\frac{\partial E}{\partial t}=\frac{I}{I_0}=\frac{dg}{dt}=h(t). \quad (3.14)$$

$h(t)$ is a function of t alone, as was mentioned previously. Thus

$$E=\int_0^t h(t')dt'+\int_0^{x^0}\rho^0(\eta)d\eta+E_0 \quad (3.15)$$

for the diode. Equation (3.12) may be integrated directly. One has

$$\frac{\kappa^3}{6}v=\int_0^t\int_0^{t'}h(t'')dt''dt'+t\left(\int_0^{x^0}\rho^0(\eta)d\eta+E_0\right)+\frac{\kappa^3}{6}v^0(x^0), \quad (3.16)$$

where $v^0(x^0)=v(x^0,0)$, and

$$\frac{\kappa^3}{6}x=\int_0^t\int_0^{t'}\int_0^{t''}h(t''')dt'''dt''dt'+\frac{t^2}{2}\left(\int_0^{x^0}\rho^0(\eta)d\eta+E_0\right)+\left[tv^0(x^0)+x^0\right]\frac{\kappa^3}{6}. \quad (3.17)$$

ρ may be obtained from Eqs. (3.17) and (3.10) if desired.

The above solution was obtained under the assumption that M is non-singular. However those electrons which were at the cathode, $x^0=0$, at $t=0$ have orbits which differ; $x^0=0$ is a singular point, therefore, and the solution is inapplicable for electrons emitted when $t>0$. Equations (3.15) to (3.17) are valid only when $0<x^0\leq 1$. The orbit of the electron emitted at $t=0$ can be obtained by a limiting operation:

$$x(t)=\lim_{x^0\to 0}x(x^0,t). \quad (3.18)$$

IV. THE STEADY STATE CASE

Those electrons which are emitted from the cathode for $t>0$ were not under consideration in the previous section but will be studied now. The transformation given by Eqs. (3.1) and (3.2) proved to be singular at the cathode, and it is necessary to use some other variable rather than the initial positions of the electrons, to distinguish them. If the dimensionless variable ξ is chosen, corresponding to the time when an electron is emitted from the cathode, then any electron's orbit is described by

$$x=x(\xi,t). \quad (4.1)$$

One uses the equation

$$t=t \quad (4.2)$$

again, and the pair of equations, (4.1) and (4.2), yield an analogous transformation to Eqs. (3.1) and (3.2). All the preceding formulas apply when x^0 is replaced by ξ.

Equation (3.15) becomes

$$E(\xi,t)=\int_\xi^t h(t')dt'+E_0(\xi), \quad (4.3)$$

where $E_0(\xi)$ is the electric field intensity at the cathode ($t=\xi$). Similarly

$$\frac{\kappa^3}{6}v(\xi,t)=\int_\xi^t\int_\xi^{t'}h(t'')dt''dt'+(t-\xi)E_0(\xi)+v_0(\xi)\kappa^3/6=\int_\xi^t(t-t')h(t')dt'+(t-\xi)E_0(\xi)+v_0(\xi)\kappa^3/6 \quad (4.4)$$

is the solution for $v(\xi,t)$ when $v_0(\xi)$ is the velocity with which electrons are emitted from the cathode.

The orbit is given by

$$\frac{\kappa^3}{6}x(\xi,t)=\int_\xi^t\int_\xi^{t'}\int_\xi^{t''}h(t''')dt'''dt''dt'+\frac{(t-\xi)^2}{2}E_0(\xi)+(t-\xi)v_0(\xi)\kappa^3/6=\int_\xi^t\frac{(t-t')^2}{2}h(t')dt'+\frac{(t-\xi)^2}{2}E_0(\xi)+(t-\xi)v_0(\xi)\kappa^3/6 \quad (4.5)$$

since $x(\xi,\xi)=0$; namely, this corresponds to the cathode.

The matrix of the transformation given by Eqs. (4.1) and (4.2) is non-singular if and only if

$$\partial x/\partial\xi\neq 0. \quad (4.6)$$

Equations (4.6) and (4.5) lead to the condition† that

$$-\frac{(t-\xi)^2}{2}\left[h(\xi)-\frac{dE_0(\xi)}{d\xi}\right]+(t-\xi)\left[\frac{\kappa^3}{6}\frac{dv_0(\xi)}{d\xi}-E_0(\xi)\right]-\frac{\kappa^3}{6}v_0(\xi)\neq 0 \quad (4.7)$$

for all t and ξ with $t\neq\xi$, since at the cathode all orbits intersect. If $E_0(\xi)$ and $v_0(\xi)$ are constant in time then (4.7) becomes

$$\frac{\kappa^3}{6}v_0(\xi)+\frac{(t-\xi)^2}{2}h(\xi)+(t-\xi)E_0(\xi)\neq 0. \quad (4.8)$$

† J. Müller (reference 2) has obtained this as a sufficient condition by different methods.

The subsequent discussion will apply only to the space charge limited diode, for which

$$E_0(\xi)=v_0(\xi)\equiv 0. \tag{4.9}$$

One has then that orbits will not intersect away from the cathode ($t\neq\xi$) if and only if

$$I(\xi)\neq 0. \tag{4.10}$$

Whenever the current density is represented by Eq. (2.7) then it follows from (4.10) that

$$\mu<1 \tag{4.11}$$

is the necessary and sufficient condition that any two electron orbits shall not have a point in common in the space between cathode and plate.

The inequality (4.11) will be assumed throughout the remainder of this paper.

The space charge density distribution for the space charge limited diode is, from Eq. (2.13),

$$\rho(\xi,t)=\frac{\partial E}{\partial x}=\frac{\partial E}{\partial \xi}\Big/\frac{\partial x}{\partial \xi}=-h(\xi)\Big/\frac{\partial x}{\partial \xi}, \tag{4.12}$$

or

$$\rho(\xi,t)=\frac{2\kappa^3}{6(t-\xi)^2}. \tag{4.13}$$

Whenever ξ is known as a function of x and t, then ρ can be given in terms of these variables. The dependence of ξ on x and t is obtained from the inversion of Eq. (4.5), an operation which is feasible, generally, by using graphical or numerical methods.

V. TRANSIT TIME

The time which is required for an electron to travel the distance between cathode and plate is defined as the transit time. Let ξ_1 be the time at which an electron emerges from the cathode, and t the time at which it arrives at the plate; then

$$\frac{\kappa^3}{6}=\int_{\xi_1}^{t}\frac{(t-t')^2}{2}h(t')dt', \tag{5.1}$$

since $x=1$ is the coordinate of the plate, by Eq. (2.5). Equation (5.1) determines ξ_1, when $h(t)$ is known.

By definition the transit time T is

$$T=t-\xi_1. \tag{5.2}$$

If $h(t)$ is of the form given by Eq. (2.7) then Eq. (5.1) becomes

$$\kappa^3=(t-\xi_1)^3+3\mu\int_{\xi_1}^{t}(t-t')^2f(t')dt'. \tag{5.3}$$

Furthermore, if one considers the static diode, in which the current is constant, then Eq. (5.3) reduces to the Child-Langmuir solution

$$T=\kappa. \tag{5.4}$$

It is necessary to invert Eq. (5.1) for ξ_1 in order that T be known explicitly. One important property of T may be deduced without performing the inversion: if $h(t)$ is periodic then T will be periodic and of the same period.

The dimensionless period is unity and therefore

$$\begin{aligned}\frac{\kappa^3}{6}&=\int_{\xi_1(t+1)}^{t+1}\frac{(1+t-t')^2}{2}h(t')dt'\\&=\int_{\xi_1(t+1)-1}^{t}\frac{(t-t'')^2}{2}h(t''+1)dt''\\&=\int_{\xi_1(t+1)-1}^{t}\frac{(t-t'')^2}{2}h(t'')dt''\\&=\int_{\xi_1(t)}^{t}\frac{(t-t')^2}{2}h(t')dt'.\end{aligned} \tag{5.5}$$

It follows that

$$T(t+1)=t+1-\xi_1(t+1)=t-\xi_1(t)=T(t). \tag{5.6}$$

ξ_1, and hence T, may be obtained explicitly when μ is small. One expands ξ_1 in a MacLaurin series

$$\xi_1=\xi_0+\mu\left(\frac{d\xi_1}{d\mu}\right)_0+\frac{\mu^2}{2!}\left(\frac{d^2\xi_1}{d\mu^2}\right)_0+\cdots, \tag{5.7}$$

where the derivatives are evaluated at $\mu=0$, as is denoted by the subscript zero.

$$\xi_0=t-\kappa. \tag{5.8}$$

All the derivatives may be obtained from Eq. (5.3). For example, the first derivative is

$$\frac{d\xi_1}{d\mu}=\frac{\int_{\xi_1}^{t}(t-t')^2f(t')dt'}{(t-\xi_1)^2[1+\mu f(\xi_1)]} \tag{5.9}$$

with

$$\left(\frac{d\xi_1}{d\mu}\right)_0=\frac{\int_{\xi_0}^{t}(t-t')^2f(t')dt'}{\kappa^2}. \tag{5.10}$$

VI. THE PLATE VOLTAGE

At any fixed instant of time an electric field intensity exists at every point between cathode and plate. Minus one times the integral of this intensity with respect to distance, between two points, is defined as the voltage difference between these points for the fixed time under

consideration. Thus the plate voltage is, by definition,

$$V'(t')=-\int_0^a E'(x', t')dx'$$
$$=-\frac{a\tau I_0}{\epsilon_0}\int_0^1 E(x, t)dx=\frac{a\tau I_0}{\epsilon_0}V(t), \quad (6.1)$$

where the integrals are evaluated for the fixed time, t.

The dimensionless plate voltage, $V(t)$, may be written as

$$V(t)=\frac{3}{\kappa^3}\int_{\xi_1(t)}^t E(\xi, t)(t-\xi)^2h(\xi)d\xi. \quad (6.2)$$

As before, $V(t)$ will be periodic with the same period as $h(t)$, if $h(t)$ is periodic. One has

$$V(t+1)=\frac{3}{\kappa^3}\int_{\xi_1(t+1)}^{t+1} E(\xi, t+1)(t+1-\xi)^2h(\xi)d\xi$$
$$=\frac{3}{\kappa^3}\int_{\xi_1(t+1)-1}^t E(\eta+1, t+1)$$
$$\times(t-\eta)^2h(\eta+1)d\eta=V(t). \quad (6.3)$$

Use has been made of the periodicity of the transit time, and of the equation

$$E(\xi+1, t+1)=E(\xi, t), \quad (6.4)$$

which can be verified readily by referring to Eq. (4.3) and noting the $E_0(\xi)=0$.

Equation (6.2) is an expression for the voltage in terms of the known current density. The inverse relation is desired in many problems, that is, the dependence of the current on the voltage. Unfortunately the inversion of Eq. (6.2) involves the inversion of Eq. (5.1), which presents considerable analytical difficulty if the current density is an arbitrary function of time. These difficulties may disappear under some circumstances, but this problem will not be treated. The current density is assumed to be known.

The MacLaurin expansion for $V(t)$ is

$$V(t)=V_0+\mu\left(\frac{dV}{d\mu}\right)_0+\frac{\mu^2}{2!}\left(\frac{d^2V}{d\mu^2}\right)_0+\cdots \quad (6.5)$$

with

$$V_0=\frac{3}{\kappa^3}\int_{\xi_0}^t(t-\xi)^3d\xi=\frac{3\kappa}{4}. \quad (6.6)$$

The subscript notation is the one employed before, in Section V. The first derivative can be calculated by using Eqs. (6.2) and (5.9). One obtains

$$\frac{\kappa^3}{3}\frac{dV}{d\mu}=\int_{\xi_1}^t(t-\xi)^2\left[\frac{dE}{d\mu}h(\xi)+E\frac{dh}{d\mu}\right]d\xi$$
$$-E(\xi, t)\int_{\xi_1}^t(t-\xi)^2\frac{dh}{d\mu}d\xi. \quad (6.7)$$

This may be simplified by integrating by parts twice. The result is

$$\frac{\kappa^3}{6}\frac{dV}{d\mu}=\int_{\xi_1}^t[1+\mu f(\xi)]\int_\xi^t(t-t')$$
$$\times\int_{t'}^t f(t'')dt''dt'd\xi, \quad (6.8)$$

with

$$\frac{\kappa^3}{6}\left(\frac{dV}{d\mu}\right)_0=\int_{\xi_0}^t\int_\xi^t\int_{t'}^t(t-t')f(t'')dt''dt'd\xi$$
$$=\int_{\xi_0}^t\int_\xi^t(t-t')\mathcal{E}(t', t)dt'd\xi, \quad (6.9)$$

where

$$\mathcal{E}(\xi, t)=\int_\xi^t f(t')dt'.$$

The second derivative is given ultimately by

$$\frac{\kappa^3}{6}\frac{d^2V}{d\mu^2}=\frac{2}{(t-\xi_1)^2}\left[\int_{\xi_1}^t(t-t')\int_{t'}^t f(t'')dt''\right]^2$$
$$-\int_{\xi_1}^t(t-\xi)\mathcal{E}^2(\xi, t)d\xi, \quad (6.10)$$

with

$$\frac{\kappa^3}{6}\left(\frac{d^2V}{d\mu^2}\right)_0=\frac{2}{\kappa^2}\left[\int_{\xi_0}^t(t-\xi)\mathcal{E}(\xi, t)d\xi\right]^2$$
$$-\int_{\xi_0}^t(t-\xi)\mathcal{E}^2(\xi, t)d\xi. \quad (6.11)$$

Higher derivatives may be obtained in a similar fashion.

VII. FOURIER EXPANSIONS

Fourier expansions are given in this section for the electric field intensity, the position of an electron, various derivatives of the voltage, and for an auxiliary function of some usefulness in later discussions, when the Fourier series for the current density is known.

It should be remarked first that there is no loss in generality in taking $f(t)$ a periodic function of period one and zero mean. One may write then

$$f(t)=\sum_{-\infty}^{\infty}f_ne^{i\omega_nt}=\sum_{n=1}^{\infty}a_n\cos\omega_nt+b_n\sin\omega_nt \quad (7.1)$$

with $\omega_n=2\pi n$ and $f_0=0$. Since $f(t)$ is a real function of the real variable t, then f_n and f_{-n} are conjugate complex quantities, and the a_n and b_n are real.

Equation (4.3) becomes

$$\mathcal{E}(\xi, t)=\sum_{-\infty}^{\infty}\frac{f_n}{i\omega_n}(e^{i\omega_nt}-e^{i\omega_n\xi}). \quad (7.2)$$

The position x is given by

$$\frac{\kappa^3}{6}x(\xi,t)=\frac{(t-\xi)^3}{6}+\mu\sum_{-\infty}^{\infty}\frac{if_n}{\omega_n}e^{i\omega_n t}\left[\frac{(t-\xi)^2}{2}e^{-i\omega_n(t-\xi)}-\frac{i}{\omega_n}(t-\xi)e^{-i\omega_n(t-\xi)}+\frac{1}{\omega_n^2}(1-e^{-i\omega_n(t-\xi)})\right]. \quad (7.3)$$

It is convenient to introduce the function $G(\xi, t)$, defined by

$$G(\xi,t)=\int_\xi^t\int_{t'}^t(t-t')f(t'')dt''dt'=\int_\xi^t(t-t')\mathcal{S}(t',t)dt'. \quad (7.4)$$

Its Fourier series is

$$G(\xi,t)=\sum_{-\infty}^{\infty}-\frac{if_n}{\omega_n}e^{i\omega_n t}\left[\frac{(t-\xi)^2}{2}-\frac{i(t-\xi)}{\omega_n}\times e^{-i\omega_n(t-\xi)}+\frac{1}{\omega_n^2}(1-e^{-i\omega_n(t-\xi)})\right] \quad (7.5)$$

with

$$G(\xi_0,t)=\sum_{-\infty}^{\infty}B_n e^{i\omega_n t}, \quad (7.6)$$

$$B_n=-\frac{if_n}{\omega_n}\left[\frac{\kappa^2}{2}-\frac{i\kappa}{\omega_n}e^{-i\omega_n\kappa}+\frac{1}{\omega_n^2}(1-e^{-i\omega_n\kappa})\right]. \quad (7.7)$$

It follows from Eqs. (7.4) and (6.9) that

$$\frac{\kappa^3}{6}\left(\frac{dV}{d\mu}\right)_0=\int_{\xi_0}^t G(\xi,t)d\xi=\sum_{-\infty}^{\infty}-\frac{if_n}{\omega_n}e^{i\omega_n t}\left[\frac{\kappa^3}{6}+\frac{\kappa}{\omega_n^2}(1+e^{-i\omega_n\kappa})+\frac{2i}{\omega_n^3}(1-e^{-i\omega_n\kappa})\right]. \quad (7.8)$$

The Fourier series for the second derivative will not be given in detail. As can be seen from Eq. (6.11) and the definition for $G(\xi, t)$ given above, the second derivative contains the square of the series for $G(\xi, t)$ and an additional series. This last expansion may be written as

$$\int_{\xi_0}^t(t-\xi)\mathcal{S}^2(\xi,t)d\xi=\sum_{-\infty}^{\infty}G_n e^{i\omega_n t}, \quad (7.9)$$

where

$$G_n=\sum_{-\infty}^{\infty}-\frac{f_r f_{n-r}}{\omega_n\omega_{n-r}}\left\{\frac{\kappa^2}{2}-i\kappa\left[\frac{e^{-i\omega_r\kappa}}{\omega_r}+\frac{e^{-i\omega_{n-r}\kappa}}{\omega_{n-r}}-\frac{e^{-i\omega_n\kappa}}{\omega_n}\right]-\left[\frac{e^{-i\omega_r\kappa}-1}{\omega_r^2}+\frac{e^{-i\omega_{n-r}\kappa}-1}{\omega_{n-r}^2}-\frac{e^{-i\omega_n\kappa}-1}{\omega_n^2}\right]\right\}. \quad (7.10)$$

VIII. AVERAGE POWER

The power exchange between the d.c. and the a.c. components is of considerable interest. Earlier workers have shown that a negative resistance may occur if the diode is operating in a proper regime,[2-4,6] and a few self-excited oscillators have been built which make use of this phenomenon.[2-3] This section will be concerned with these questions.

The general expressions given above enable one to compute power for all $0\leq\mu<1$. As has been mentioned before, the analytic treatment is difficult when μ is not small. Attention will be restricted, therefore, to a small signal analysis.

The instantaneous power per unit area P' is defined as

$$P'=V'(t')I(t') \quad (8.1)$$

or

$$P'=\frac{a\tau I_0^2}{\epsilon_0}V(t)[1+\mu f(t)]=\frac{a\tau I_0^2}{\epsilon_0}P(t). \quad (8.2)$$

The average power per unit area, $\langle P'\rangle$, is the time average of P' over a cycle, and is given by

$$\langle P'\rangle=\frac{1}{\tau}\int_0^\tau P'(t')dt'=\frac{a\tau I_0^2}{\epsilon_0}\int_0^1 P(t)dt=\frac{a\tau I_0^2}{\epsilon_0}\langle P\rangle. \quad (8.3)$$

The MacLaurin series expansion for $\langle P\rangle$ is

$$\langle P\rangle=\langle P\rangle_0+\mu\left(\frac{d\langle P\rangle}{d\mu}\right)_0+\frac{\mu^2}{2!}\left(\frac{d^2\langle P\rangle}{d\mu^2}\right)_0+\cdots, \quad (8.4)$$

with the usual subscript notation.

The use of Eqs. (8.3), (6.9), and (6.11) yields the set of equations

$$\langle P\rangle=\frac{3\kappa}{4}, \quad (8.5)$$

$$\left(\frac{d\langle P\rangle}{d\mu}\right)_0=\int_0^1\left(\frac{dV}{d\mu}\right)_0dt+\int_0^1 V_0 f(t)dt=0, \quad (8.6)$$

$$\left(\frac{d^2\langle P\rangle}{d\mu^2}\right)_0=\int_0^1\left(\frac{d^2V}{d\mu^2}\right)_0dt+2\int_0^1\left(\frac{dV}{d\mu}\right)_0 f(t)dt. \quad (8.7)$$

Equation (8.6) states that the first-order correction to the d.c. power is zero. The second-order correction con-

tains two integrals which are, as will be shown, of comparable magnitude under certain conditions of diode operation. The first integral is the contribution to the power arising from the second-order correction to the voltage; it appears to have been neglected in previous discussions of diode theory. The second integral is the only contribution to the power which has been discussed in the literature. It may be interpreted as the diode "resistance" times the mean of the square of the alternating current.

The first integral is given by

$$\int_0^1 \left(\frac{d^2V}{d\mu^2}\right)_0 dt = 24\kappa \sum_{r=1}^{\infty} \frac{f_r f_{-r}}{\theta_r^4} \times\left[1-\frac{\theta_r^2}{4}+\frac{2}{\theta_r^2}(1-\cos\theta_r)-\frac{2\theta_r\sin\theta_r}{\theta_r^2}\right], \quad (8.8)$$

and the second by

$$\int_0^1 \left(\frac{dV}{d\mu}\right)_0 f(t)dt = 12\kappa \sum_{r=1}^{\infty} \frac{f_r f_{-r}}{\theta_r^4} \times[2(1-\cos\theta_r)-\theta_r\sin\theta_r], \quad (8.9)$$

where $\theta_r=\omega_r\kappa$.

Substituting Eqs. (8.8) and (8.9) into Eq. (8.7), one obtains

$$\left(\frac{d^2\langle P\rangle}{d\mu^2}\right)_0 = 24\kappa \sum_{r=1}^{\infty} \frac{f_r f_{-r}}{\theta_r^4}\left[1-\theta_r\sin\theta_r\left(1+\frac{2}{\theta_r^2}\right)+2(1-\cos\theta_r)\left(1+\frac{1}{\theta_r^2}\right)-\frac{\theta_r^2}{4}\right]. \quad (8.10)$$

The dimensionless quantity $\langle P\rangle$ is, therefore, to second-order

$$\langle P\rangle = \frac{3\kappa}{4}+12\mu^2\kappa \sum_{r=1}^{\infty} \frac{f_r f_{-r}}{\theta_r^4}\left[1-\theta_r\sin\theta_r\left(1+\frac{2}{\theta_r^2}\right)+2(1-\cos\theta_r)\left(1+\frac{1}{\theta_r^2}\right)-\frac{\theta_r^2}{4}\right], \quad (8.11)$$

or the average power per unit area is

$$\langle P'\rangle = \left(\frac{6a^4 I_0^5}{\epsilon_0^2}\frac{m}{e}\right)^{\frac{1}{2}}\left\{\tfrac{3}{4}+12\mu^2\sum_{r=1}^{\infty}\frac{f_r f_{-r}}{\theta_r^4} \times\left[1-\theta_r\sin\theta_r\left(1+\frac{2}{\theta_r^2}\right)+2(1-\cos\theta_r)\left(1+\frac{1}{\theta_r^2}\right)-\frac{\theta_r^2}{4}\right]\right. \quad (8.12)$$

Llewellyn[3] has treated the case $f_r=0$, $r\neq 1$ and obtained the reduced form of Eq. (8.10), aside from constants, as the expression for the "resistance" of the diode. However, Eq. (6.11) shows that the "resistance" (the real part of the complex impedance at a single frequency) times the mean of the square of the single frequency alternating current does not give the power at this frequency. The additional terms are given by Eq. (8.8).

For a given frequency the ratio λ of the total second-order correction to the power to the "resistance" term is given by one of (8.12) to one of (8.10). Thus,

$$\lambda = \frac{1-\theta_r\sin\theta_r[1+(2/\theta_r^2)]+2(1-\cos\theta_r)[1+(1/\theta_r^2)]-\theta_r^2/4}{2(1-\cos\theta_r)-\theta_r\sin\theta_r}. \quad (8.13)$$

For small values of θ_r this becomes

$$\lambda = \tfrac{5}{6}+O(\theta_r^2) \quad (8.14)$$

and the large values of θ_r

$$\lambda = O(\theta_r), \quad (8.15)$$

where $O(x)$ refers to the usual order of greatness.[7] Thus the terms neglected in the Llewellyn treatment dominate the "resistance" term in the expression for the power for large θ_r. Even for small θ_r the correction is about 16 percent.

[7] E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge University Press, London, England, 1935), p. 11.

High Frequency Impedance of Low Pressure Gaseous Diodes

CHAI YEH
*Department of Electrical Engineering, University of Kansas, Lawrence, Kansas**

E. L. CHAFFEE
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received March 20, 1950)

A simple theory of the lagging effect of the positive ions in neutralizing the space charge near the cathode of a gaseous diode is developed. By means of bridge measurement up to a frequency of 350 or 1000 kc per second, depending upon the type of tube used, the theory is checked experimentally in terms of the high frequency impedance of the diodes. Several constants of interest, such as the transit time and the lifetime of a positive ion, its effectiveness in neutralizing space charge, etc., can be deduced from the theoretical and experimental investigations.

I. INTRODUCTION

WHEN a conducting path of electricity possesses quantities which cause a lag and lead of an alternating current with respect to the voltage, the effect is ordinarily expressed in terms of inductive or capacitive reactance. This characteristic of a conductor may result from the inertia of the carriers of the charge wholly irrespective of any ordinary inductance or capacitance which the circuit may contain. Llewellyn[1] has treated the effects of the inertia of the electron at very high frequencies in modifying the conduction characteristics of high vacuum tubes. A similar inertia effect may be observed at low frequencies in the conduction through a gaseous vacuum tube. The reactance of the conduction path in this case results from the inertia of the positive ions. An interesting case of this type of conduction with inertia effects is described in this paper. A simple theory of the effect will first be presented and then experimental measurements on several soft tubes will be given to show that the theory at least approximately describes the effects observed. Certain fundamental quantities pertaining to the conduction process can be approximately determined from the experimental results.

The case to be treated is that of a cylindrical diode containing a trace of gas. The d.c. polarizing potential is adjusted to be slightly greater than the ionization potential of the gas and a small superimposed a.c. voltage is applied. The object is to find the equivalent a.c. resistance and reactance of the diode as a function of frequency.

The conditions assumed in the analysis are as follows: The static characteristic curve of the soft diode is of the type shown in Fig. 1, where $\bar{E}_i$ is the ionization potential. The increase in current above the normal space-charge limited current when the impressed potential exceeds $\bar{E}_i$ is assumed to be due to the partial neutralization of space charge near the cathode by positive ions. The current carried by positive and negative ions is assumed negligible.

The instantaneous anode current i at time t is not only a function of the instantaneous anode potential e but is also a function of the applied anode potential at a previous time $t-\tau$ denoted by $e_{t-\tau}$. The reason for the latter effect is that positive ions produced near the anode at time $t-\tau$ requires time τ to reach the space near the cathode where they neutralize some of the negative space charge. Expressed analytically,

$$i=f(e, e_{t-\tau}). \tag{1}$$

Equation (1) can be expressed in terms of n, the number of positive ions existing at time t in the cathode space. Then

$$i=F(e, n). \tag{2}$$

Since n does not depend upon e but is a function of $e_{t-\tau}$, e and n can be considered as independent variables. Then the differential of Eq. (2) is

$$di=(\partial i/\partial e)_n de+(\partial i/\partial n)_e dn=Kde+Gdn, \tag{3}$$

where $K\equiv(\partial i/\partial e)_n$, and is the slope of the $i-e$ curve for a constant number of positive ions, and $G\equiv(\partial i/\partial n)_e$, and is the effectiveness of partial neutralization of space charge by the positive ions at constant e.

II. CALCULATION OF THE NUMBER OF POSITIVE IONS

The evaluation of the current by Eq. (3) requires a knowledge of the number of positive ions, n at time t.

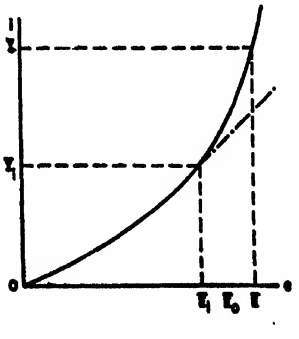


FIG. 1. Static characteristic curve of a gaseous diode. The dot-dash extension indicates the space-charge limited case in a high vacuum tube.

* Formerly, Cruft Laboratory, Harvard University, Cambridge, Massachusetts.

[1] F. B. Llewellyn, Proc. I. R. E. 21, 1532 (1933), 22, 947 (1934), 23, 112 (1935).

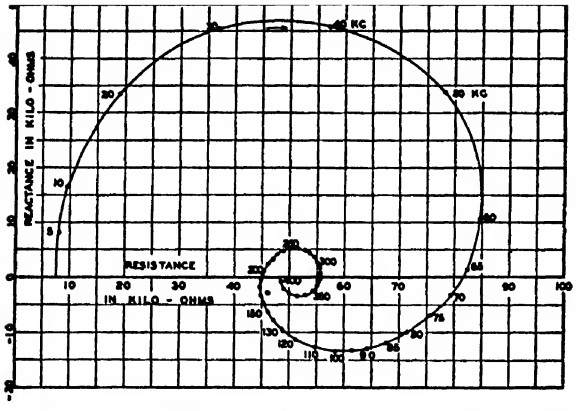


FIG. 2. Theoretical impedance curve of Type I, $B<K$, calculated with the following constants: $\bar{I}=0.2$ ma, $\bar{E}_0=1$ volt, $K=0.1955\cdot10^{-4}$ mhos, $P=0.1445\cdot10^{6}$ 1/volt²-sec., $\beta=0.401\cdot10^{5}$ 1/sec., $\tau=3.960\cdot10^{-6}$ sec. Frequency scale is spotted along the curve.

An expression for n will now be derived. Assume a cylindrical diode having an anode of radius a at potential e. The diode is assumed to contain a trace of gas which has an ionization potential $\bar{E}_i$. If $\bar{E}_i$ is slightly less than e, ionization occurs in a thin cylindrical space adjacent to the anode. The gas pressure is such that the mean free path of electrons is large compared with a.

Let $S(e_r)$ be the ionization coefficient at potential e_r, defined as the number of ions produced by one electron in traveling one centimeter at one-mm pressure. Then the number of ions produced by current i in time dt is

$$dN=(pi/10\epsilon)dt\int_{r_i}^{a}S(e_r)dr, \tag{4}$$

where p is the pressure in mm of mercury, i is the current in amperes, ϵ is the electronic charge in absolute e.m.u., and r_i is the radius corresponding to $\bar{E}_i$.

We shall assume a linear relation between e_r and r near the anode in the form $e_r=[(r-a)/g-1]e$, so that $dr=(g/e)de_r$. Furthermore, we shall assume a linear relation between $S(e_r)$ and e_r of the form $S(e_r)=b(e_r-\bar{E}_i)$. This latter relation is quite accurate for values of e_r not exceeding $2\bar{E}_i$. Using these relations, Eq. (4) becomes

$$dN=M(i/e)\cdot(e-\bar{E}_i)^2dt, \tag{5}$$

where

$$M=pgb/20\epsilon. \tag{6}$$

M is a constant for any given tube at constant pressure.

Equation (5) gives the number of positive ions produced in time dt near the anode. We shall indicate the mean transit time of a positive ion in traveling from anode to cathode by τ. Although some positive ions may strike the cathode, others miss the cathode and oscillate about it. We must therefore recognize a mean lifetime of positive ions in the cathode space. The decay of any elementary group dN in time t when near the cathode may be expressed by the relation

$$dn=dNe^{-\beta t}, \tag{7}$$

where β is the reciprocal of the mean lifetime of a positive ion as defined by Kingdon.[3] Multiplying Eq. (5) by (7) in which $t-\tau-\lambda$ is substituted for t, we have for the number of positive ions at time t,

$$n=\int_{-\infty}^{t-\tau}M(e_\lambda-\bar{E}_i)^2(i_\lambda/e_\lambda)e^{-\beta(t-\tau-\lambda)}d\lambda. \tag{8}$$

Now assume that $\bar{E}$ and $\bar{I}$ are the d.c. anode potential and anode current, where $\bar{E}$ is assumed to be just greater than $\bar{E}_i$. Let a small a.c. potential of angular frequency ω be superimposed on $\bar{E}$. Then

$$\begin{aligned}e&=\bar{E}+\Delta E\sin\omega t,\\ i&=\bar{I}+\Delta I\sin(\omega t-\phi).\end{aligned} \tag{9}$$

Substituting expressions (9) for e_λ and i_λ in (8) and

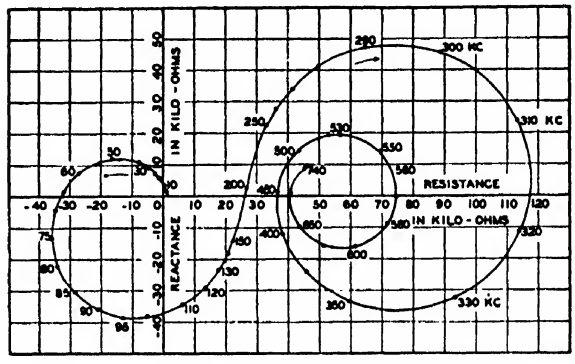


FIG. 3. Theoretical impedance curve of Type II, $B>K$, calculated with the same set of constants as the Type I curve except with $\bar{I}=1.5$ ma. The constants used in these calculations were actually determined from experimental data. The results check with each other very well up to 150 kc. Measurements at higher frequencies are lacking due to experimental difficulties.

integrating, we obtain n in the form $N+\Delta n$ plus high order terms. Retaining terms of fundamental frequency only and neglecting terms of the higher order of ΔE and ΔI, gives

$$\Delta n=\frac{\bar{E}_0M}{\bar{E}(\omega^2+\beta^2)^{\frac{1}{2}}}\{\Delta E\bar{I}(2-\bar{E}_0/\bar{E})\sin[\omega(t-\tau)-\theta]+\bar{E}_0\Delta I\cos[\omega(t-\tau)-\theta-\phi]\}, \tag{10}$$

where $\bar{E}_0=\bar{E}-\bar{E}_i$, $\theta=\tan^{-1}\omega/\beta$.

III. ALTERNATING CURRENT IMPEDANCE

We can now substitute (10) in (3) and equate to zero the coefficients of $\sin\omega t$ and $\cos\omega t$. Since $\sin\phi=X/Z$ and $\cos\phi=R/Z$, the values of R and X can be found and are

$$R=\frac{(K-BA)+(B-AK)\cos\delta}{K^2+B^2+2KB\cos\delta}, \tag{11}$$

$$X=\frac{(B+AK)\sin\delta}{K^2+B^2+2KB\cos\delta}, \tag{12}$$

[3] K. H. Kingdon, Phys. Rev. 21, 408 (1923).

$$Z^2=R^2+X^2=\frac{A^2-2A\cos\delta+1}{K^2+B^2+2KB\cos\delta}, \tag{13}$$

where

$$A=\frac{MG\bar{E}_0^2}{\bar{E}(\omega^2+\beta^2)^{\frac{1}{2}}}=\frac{P\bar{E}_0^2}{(\omega^2+\beta^2)^{\frac{1}{2}}}, \quad P=\frac{MG}{\bar{E}}, \tag{14}$$

$$B=\frac{MG\bar{E}_0\bar{I}}{\bar{E}(\omega^2+\beta^2)^{\frac{1}{2}}}\left(2-\frac{\bar{E}_0}{\bar{E}}\right)=\frac{P\bar{E}_0\bar{I}}{(\omega^2+\beta^2)^{\frac{1}{2}}}\left(2-\frac{\bar{E}_0}{\bar{E}}\right) \tag{15}$$

and

$$\delta=\omega\tau+\theta=\omega\tau+\tan^{-1}\omega/\beta. \tag{16}$$

It is interesting first to investigate the resistance and reactance at the two extreme frequencies, that is at zero and infinite frequencies. Substituting $\omega=0$, we have

$$\begin{aligned} X_0&=0,\\ \phi_0&=0,\\ R_0&=(1-A)/(K+B), \end{aligned} \tag{17}$$

where the zero subscript indicates the value at zero frequency. Here we find that the reactance is zero and the resistance is finite and positive as long as A is less than unity, which is true in practice. And for $\omega=\infty$, we have

$$\begin{aligned} X_\infty&=0,\\ \phi_\infty&=0,\\ R_\infty&=Z_\infty=1/K. \end{aligned} \tag{18}$$

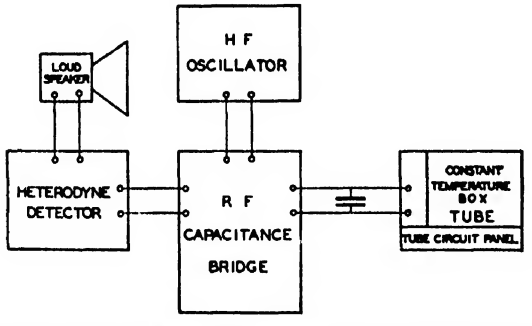


FIG. 4. Experimental set up for the measurement of high frequency impedance of a gaseous diode.

IV. THEORETICAL IMPEDANCE CURVES

Equations (11) and (12) represent the resistance and reactance components of the complex impedance of a gaseous diode. From Eqs. (14)–(16), it is seen that both the constants A and B and the transit angle δ are functions of frequency. An investigation of the frequency characteristics of this complex impedance can be made either by plotting R and X individually as a function of frequency, or by plotting the reactance against resistance. For the present discussion, the latter presentation is preferrable. As an aid in determining the shape of the impedance curve given by Eqs. (11) and (12), the intercepts on the R and X axis may first be found.

The intercepts on the X or vertical axis can be found by letting $R=0$, and solving Eqs. (11) and (12) simultaneously. The result is

$$X=\pm\left(\frac{1-A^2}{B^2-K^2}\right)^{\frac{1}{2}}, \ (R=0). \tag{19}$$

As A is always less than unity (which is required in order to have positive resistance at zero frequency), then $1-A^2$ is a positive quantity. If B^2 is greater than K^2, the term under the square root sign is real, and the positive and negative signs in front of it indicate the positive and negative intercepts on the X-axis corresponding to different frequencies. But if B^2 is less than K^2, the term under the square root sign becomes imaginary, which means that there are no intercepts.

The intercepts on the R or horizontal axis can be found by letting $X=0$ in Eqs. (11) and (12). Thus we have

$$R=(1+A)/(K-B), \quad (X=0). \tag{20}$$

It is seen that R can be either positive or negative according to whether K is greater or less than B, since both B and K are positive quantities. The above analysis leads to the discussion of two types of the impedance curves. They are very distinct in shape and are characterized by whether $B<K$ of $B>K$.

Type I, $B<K$.—In this case, the entire curve lies on the positive resistance side of the diagram, and there is no intercept on the X-axis. At low frequencies, the resistance is nearly constant while the inductive reactance increases with frequency. Above a certain frequency, both resistance and reactance change rapidly. The reactance changes from inductive to capacitive as $\sin\delta$ changes its sign. The resistance swings back and forth about the value of R at $\omega=\infty$ but remains positive. Both the resistance and reactance decrease in magnitude as ω increases and the curve spirals in a clockwise direction in towards the point corresponding to $1/K$ at $\omega=\infty$. This is indicated in Fig. 2, which is a plot of the theoretical curve of Type I for the following constants:

$\bar{I}=0.2$ ma $\quad \bar{E}_0=1$ volt, $\quad K=0.1955\cdot10^{-4}$ mhos,
$P=0.1445\cdot10^{5}$ 1/volt²-sec., $\quad \beta=0.401\cdot10^{5}$ 1/sec.,
$\tau=3.96\cdot10^{-6}$ sec.

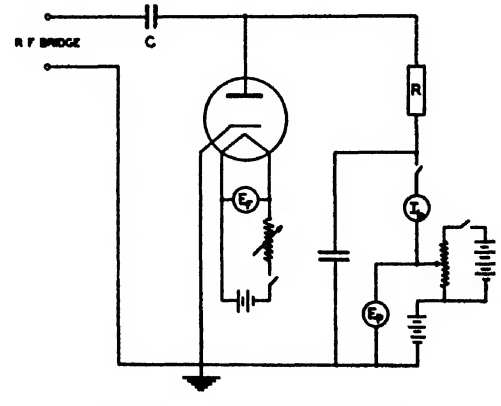


FIG. 5. Experimental tube circuit.

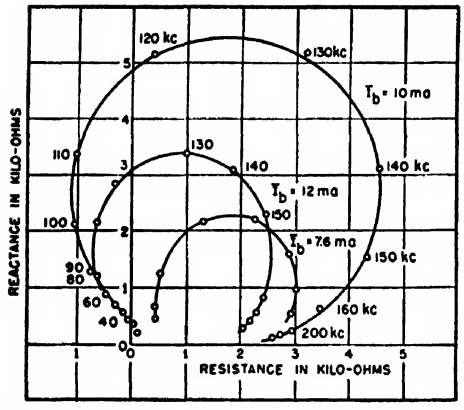


FIG. 6. Experimental impedance curve of a helium-filled diode.

Type II, for $B>K$.—For this type of curve, there are two intercepts on the X-axis which are $X=+[(1-A^2)/(B^2-K^2)]^{\frac{1}{2}}$ and $-[(1-A^2)/(B^2-K^2)]^{\frac{1}{2}}$ corresponding

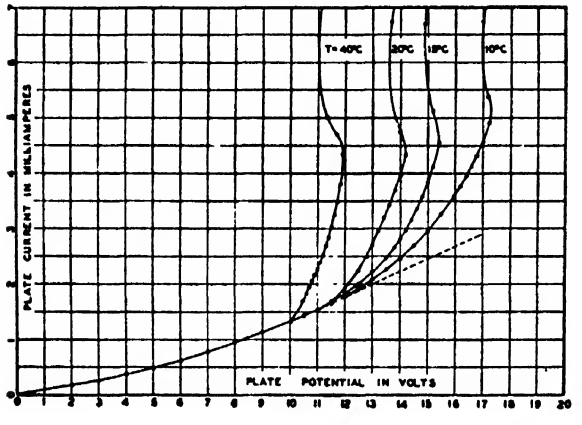


FIG. 7. Static characteristic curve of the experimental tube RK-100, containing mercury vapor.

to different frequencies. There is a negative intercept on the R-axis, which corresponds to a negative resistance. As K is a constant and a positive quantity, while B is a positive number but diminishes in magnitude as frequency increases, it is obvious that as the frequency increase further, B decreases and finally becomes less than K. Then the curve takes the form of Type I.

At very low frequencies, the resistance is positive and practically constant. As the frequency becomes greater, it decreases and becomes negative at certain frequencies and then returns to positive values progressing in a counter-clockwise direction along a spiral. For higher frequencies, an inflection point is reached and the curve spirals in a clockwise direction and the resistance varies in the manner of Type I. The entire curve of Type II is plotted in Fig. 3 for the same set of constants used for Type I curve except that $\bar{I}=1.5$ ma instead of 0.2 ma.

V. EXPERIMENT

In the theory developed so far, we made use of four unknown constants which may be determined experimentally from the intercepts of the experimental curves. These are τ, β, K, and $P(=MG/\bar{E})$. It is better to use P instead of A and B, since both A and B are derivable from P.

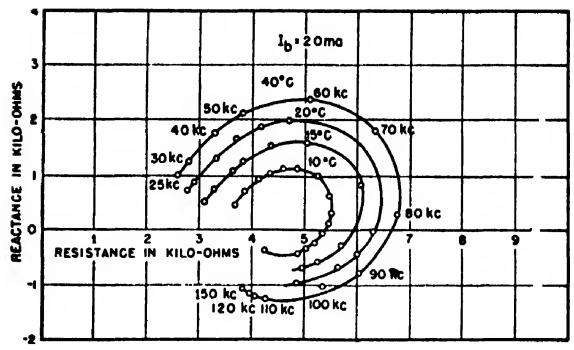


FIG. 8. Effect of gas pressure on the impedance curve at low anode current. Tube contains mercury vapor.

The general arrangement of the apparatus used for measuring the tube impedance is shown in Fig. 4. The entire set up consists of a high frequency bridge, a high frequency oscillator, a sensitive detector, and the tube circuit. The tube, whose impedance characteristic is to be measured, is mounted in a constant-temperature oil bath controlled by a thermostat. The complete tube circuit is shown in Fig. 5. In this circuit, C is a blocking condenser so as to isolate the d.c. anode potential from the bridge circuit. R is one of the several non-inductive resistances having ranges from several hundred to several hundred thousand ohms. The anode potential can be varied by battery taps and more precisely by the potential divider. The whole system is enclosed in a well-grounded aluminum box.

Before taking any measurement on the high frequency bridge, a d.c. static characteristic curve of the tube is taken. The compensation method is adopted for the measurement on the high frequency bridge. The anode potential of the tube to be measured is adjusted to a proper point on the d.c. characteristic. The output of the oscillator is kept low in order to conform to the assumption that the a.c. component is very much smaller than the d.c. values. For each frequency, the bridge is balanced by adjusting both the air condenser and the parallel resistance on the bridge. Two sets of readings are necessary; one with the anode potential of the tube on, and the other with it off. The first set of readings gives the tube admittance in parallel with that of the circuit, while the second set of readings gives the circuit admittance alone. The observed change in resistance and capacitance is that due to the tube impedance. The resistance and reactance thus measured are plotted and the intercepts on the real and imaginary axes are determined. Equations (17), (19), and (20) may then be used to evaluate the constants P, β, and K and Eq. (16) is used to calculate τ. A curve of Type II is required to provide enough intercepts to determine the four constants.

VI. RESULTS AND DISCUSSIONS

Measurements on several gaseous tubes were made. Some of the tubes were filled with helium gas, others with mercury vapor. The results obtained from the helium-filled tubes indicate that the time of flight of the positive ions is extremely small. A frequency range of from 30 to 1000 kc, is used in this experiment. The impedance curve thus obtained is plotted in Fig. 6. It is noticed that the reactance of this tube remains inductive even at extremely high frequencies. The resistance is negative during a small portion of the frequency range. But as the anode current is varied, both the resistance and reactance remain positive. The impedance curve of this tube is essentially of Type I, but has characteristics of the transition type between those of I and II where there are two positive values of X for zero values of R and where the spiraling is always clockwise. We were not able to increase the anode current $\bar{I}$ or the gas pressure sufficiently to obtain the anticlockwise spiral of Type II. The anode current saturates at a slightly higher operating point. It is therefore impossible to determine the constants of this tube.

The next group of tubes were filled with mercury vapor. Several special diodes were constructed by the Raytheon Manufacturing Company. They are cylindrical diodes with indirect-heated cathodes. The cathode radius is 0.075 cm and the anode radius is 1 cm. The d.c. characteristic curve of one of these tubes is shown in Fig. 7. The curves are taken at different temperatures of the bulb which determines the vapor pressure inside the tube. Above an anode current of 4.5 ma, a greenish glow starts and an abrupt increase of anode current results which is accompanied by a decrease in potential as is shown in the plots. All measurements were made at an anode current well below the glow discharge.

Figures 8 and 9 are taken for the purpose of illustrating the effect of varying the gas pressure on the impedance characteristics of the tube. In Fig. 8, the anode current is held constant at 2 ma, while in Fig. 9, it is held at 4 ma. It is noticed that for low anode current, the changing of gas pressure merely changes the dimension of the impedance curve without materially affecting its position or shape. These are shown as concentric curves in Fig. 8, and are essentially of Type I. For higher anode current, the variation of the gas pressure shows the evolution of the impedance curve from Type I to Type II. Figures 10 and 11 are taken for the purpose of illustrating the effect of changes of anode current on the impedance curve of the tube. Two sets of data were taken, one for low pressure and the other for high pressure. At low pressure, the shape of the impedance curve does not change very much and the increase of the anode current merely moves the curves towards lower resistance and reactance as shown in Fig. 10. For high pressures, the evolution of the impedance curves from Type I to Type II as the anode current increases is apparent as shown in Fig. 11. This time, the factor that affects B is the anode current. Hence both the anode current and the gas pressure are responsible for the change of shape of the impedance curve. For an anode current greater than 4 ma, and the bulb temperature higher than 40°C, the hiss inside the tube becomes so

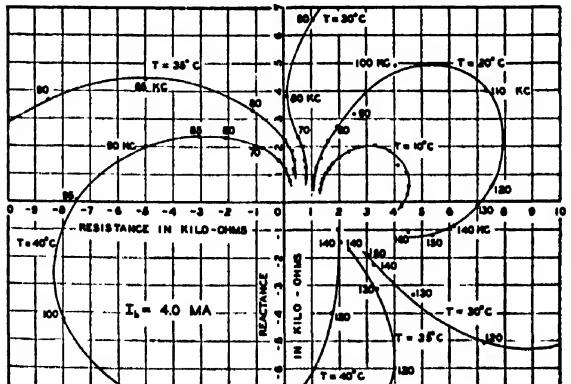


FIG. 9. Effect of gas pressure on the impedance curve at high anode current. Tube contains mercury vapor.

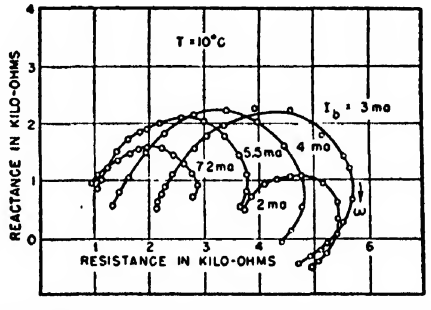


FIG. 10. Effect of anode current on the impedance curve at low gas pressure. Tube contains mercury vapor.

intense that bridge balance is difficult and inaccurate. Figure 12 is another plot of one of the curves in Fig. 11 (for $\bar{I}=4.0$ ma) together with the theoretical curve calculated from the constants preciously determined. The theoretical and experimental data are carried far into the higher frequency region in order to show the change over from an anticlockwise to clockwise rotation of the spiral as in the Type II curve.

For a typical operation corresponding to an anode current of 4.0 ma, and a mercury-vapor pressure corresponding to 40°C, the following constants were determined:

Typical operation condition	Intercepts from experimental data	Calculated constants
$\bar{E}=11.86$ volts	$R_0=350$ ohms	$K=0.239\cdot10^{-4}$ mhos
$\bar{E}_0=1.48$ volts	X (at $R=0$) $=1200$ ohms	$\beta=0.164\cdot10^{8}$ 1/sec.
$\bar{I}=4.0$ ma	R (at $X=0$) $=-7500$	$P=0.268\cdot10^{8}$ 1/V²-sec.
$T=40$°C	$\omega_1=2\pi50\cdot10^{3}$	$\tau=3.28\cdot10^{-6}$ sec.
	$\omega_2=2\pi90\cdot10^{3}$	

As an example of the application of these constants, the following computations are made. Knowing $1/\beta$, the

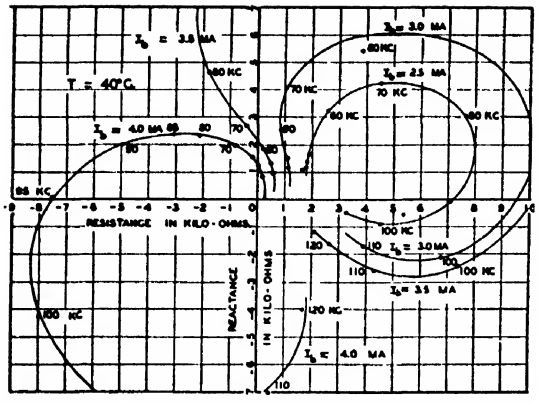


FIG. 11. Effect of anode current on the impedance curve at high gas pressure. Tube contains mercury vapor.

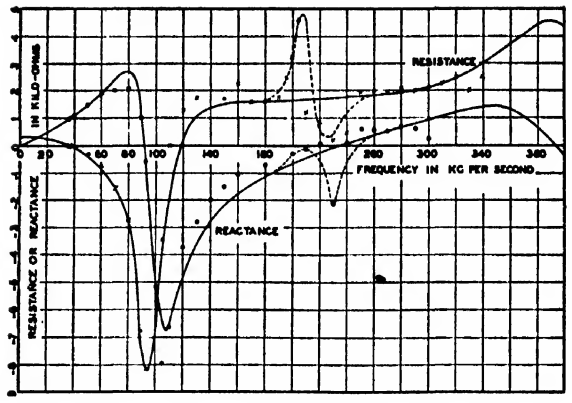


FIG. 12. Theoretical and experimental check of a typical operation: $\bar{E}=11.86$ volts, $\bar{E}_0=1.48$ volts, $\bar{I}=4.0$ ma, $T=40°$C. Theoretical curves are drawn in solid lines, the circles and crosses are the experimental values. Circles for reactance, crosses for resistance. The experiments and calculations were carried out to higher frequencies, up to 340 kc on this tube. The dotted curve indicates the region where hiss is excessive.

mean lifetime of the positive ions, we may proceed to calculate the average number of times that the positive ions circulate around the cathode provided that we know the equivalent radius of the circular path of the circulating ions. From the consideration that the centrifugal force of the moving ions are balanced by the force acting on the ions due to the electric field applied between the electrodes, and making use of γ, which is a function of r as given in Langmuir's paper,[3] the radius of the circulating path is

$$r_a=2\gamma^2/(d\gamma^2/dr_a). \tag{21}$$

This equation can best be solved by cut and try method. The solution is found for a cathode radius of 0.075 cm and is $r_a=0.080$ cm. The number of times that an ion circulates around the cathode is given by $N_a=v/2\pi r_a\beta$, where v is the velocity of the positive ions and is given by $v=(2E\epsilon/m_+)^{\frac{1}{2}}$. For a mercury ion, the average number of times that the ions circulate around the cathode for the conditions stated is 4.08 or roughly 4 times.

The constants P and β enable us to calculate approximately the average number of electrons that are neutralized by a single positive ion under the stated condition. From Eq. (7), it is seen that the rate of disappearance of positive ions is $dn/dt=-\beta N$. In order to maintain the steady state condition, the rate of supply of positive ions should be equal to βN, which is maintained by the increment in the positive ion current. If the increment in positive ions is proportional to the increment in the positive ion current, it is found that

$$\frac{\text{Number of electrons}}{\text{Number of positive ions}}=\frac{G}{\epsilon\beta}=\frac{P\bar{E}}{M\epsilon\beta}. \tag{22}$$

For an operating temperature of 313°K, which corresponds to a mercury vapor pressure of 0.004 mm of Hg, and the constant of proportionality $b=1.5$, $g=0.925$, the average number of electrons that are neutralized by a single positive ion under this operating condition is equal to 7480, or roughly 8000.

[3] I. Langmuir, Phys. Rev. 21, 435 (1923), 22, 347 (1923).

The Propagation of Plastic Deformation in Solids*

THEODORE VON KARMAN AND POL DUWEZ
Pasadena, California
(Received March 27, 1950)

The stress wave caused by a longitudinal impact at the end of a cylindrical bar has been analyzed in the case where the impact velocity is large enough to produce plastic strain. The theory gives a method for computing the stress distribution along the bar at any instant during impact. It is shown that for a given material, there is a critical impact velocity such that when subjected to a tension impact with a velocity higher than the critical, the material should break near the impacted end with negligible plastic strain.

An experimental investigation was made concurrently with the theoretical study. Some of the most significant experimental results are presented in this paper.

I. INTRODUCTION

THE testing of impact strength of materials previous to the discovery of the laws of propagation of plastic deformation used essentially two different methods: bending of notched bars, and tension or compression of short cylindrical specimens. The first method presumes a certain standardization of the shape of the specimens and tries to establish certain relative merit numbers for the impact resistance of various materials by measuring the energy absorbed by the standardized specimens. The second method is somewhat more ambitious; its aim is the determination of the energy absorption by unit volume of the material, in the case of impact, and the comparison of this amount of energy with the work absorbed in static experiments. These tests, especially the tests of the second type, have shown that for moderate impact velocities the work absorbed in impact is, in general, somewhat larger than the work absorbed in static experiments; there were some indications that this tendency is reversed as the impact velocity increases. However, the observed drop in the amount of absorbed work was not accompanied by a decrease in elongation and no explanation was offered for this phenomenon.

As the European war started, the attention of military and other engineers concerned with the construction of bomb resisting structures was focused on the impact problem. The senior author became interested in the problem of resistance of structures to impact loads beyond the elastic limit in a conference held at the National Academy of Sciences early in 1940, under the chairmanship of Dr. V. Bush. He conceived the idea of the plastic wave in order to obtain a method for theoretical calculation of the energy absorbed by various structures, as columns, beams, plates, etc. The theory of elastic wave propagation was known. It was evident from this theory that, for any especially brittle material (i.e., materials which break at their elastic limit), a critical impact velocity exists for which fracture must occur at the point of application of the impact, so that the rest of the structure has no chance to participate in the energy absorption. It appeared that if a theory of the propagation could be developed, such a theory would reveal the existence of a critical velocity for structures loaded beyond their yield strength.

The basic theory and preliminary experiments supporting the theory, at least in its fundamental conclusions, were carried out during the year 1941. The senior author presented the fundamentals of the theory in a paper sent for publication in the Proceedings of the National Academy of Sciences in December, 1941. It was found, however, that the paper might have some bearing on problems relating to National Defense. The publication was therefore delayed, but the study was described in a classified NDRC report, A-29, "On the Propagation of Plastic Deformation in Solids," February 1942. The experimental work carried out previously by the junior author was published at the same time in the classified NDRC report, A-33, "Preliminary Experiments on Propagation of Plastic Deformation." Subsequently more complete theoretical investigations on the subject were carried out by the senior author, H. F. Bohnenblust, D. H. Hyers, and J. Charyk. The main content of these investigations will probably be published in the near future.

II. LONGITUDINAL IMPACT AT THE END OF A BAR EXTENDING TO INFINITY

The propagation of elastic strains in a cylindrical bar subjected to tension impact had already been analyzed in 1807 by Thomas Young, who established the proportionality between the elastic strain and the velocity of impact. The value of the elastic strain ϵ is in this case given by the formula

$$v_1 = c_0 \epsilon,$$

in which v_1 is the velocity of impact and c_0 is the velocity of propagation of an elastic deformation. Young had pointed out in his paper that when the velocity of impact is above a certain critical value a plastic strain is initiated near the point of impact. Since Young's work, it seems that the interest of the problem has not been fully recognized, and no systematic attempt has been made to compute the stress and strain caused by an impact beyond the elastic limit of the material. In the present paper such a treatment is provided for the case of longitudinal impact.

* Presented at the Sixth International Congress for Applied Mechanics, Paris, France, September 22–29, 1946.

Consider a rod or wire extending from $x=-\infty$ to $x=0$ and assume that the endpoint at $x=0$ is suddenly put in motion with a constant velocity v_1. Let the stress-strain relation for the material be given by a function of the form $\sigma=\sigma(\epsilon)$ where σ is the stress and ϵ the strain. Stresses that depend on the time-rate of strain will be neglected; for, with the exception of the case of extremely high velocities, such stresses are small in comparison with the stresses that depend on the strain itself. To be sure, the relation $\sigma=\sigma(\epsilon)$ holds only for the first deformation of the material beyond the elastic limit; in the case of load reversal, another functional relation, which takes the hysteresis into

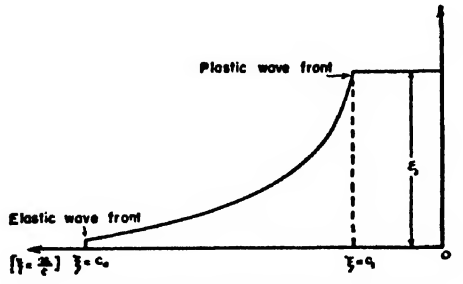


FIG. 1. Schematic representation of ϵ *versus* ξ.

account, has to be used. The lateral contraction of the material—that is, the contribution of the lateral contraction to the kinetic energy—is neglected in the following calculation.

With these simplifications, the equation of motion for an element of the rod or wire can be written in the form

$$\rho\frac{\partial^2 u}{\partial t^2}=\frac{d\sigma}{d\epsilon}\frac{\partial \epsilon}{\partial x}, \tag{1}$$

where u is the displacement of the element in the longitudinal direction, ρ is the density of the material, and t is the time.

It should be pointed out that in Eq. (1) the process of deformation is considered from the Lagrangean point of view. An arbitrary cross section of the bar is determined by a coordinate x, which gives its distance from the origin in the unloaded state. The stress is defined as the ratio of the internal reaction between two portions of the bar to the initial cross-sectional area of the bar. The strain is numerically equal to the change of length of a portion of the bar which, in the initial unloaded state, was of unit length. If u is the displacement of a cross section, the strain ϵ is equal to $\partial u/\partial x$. It is not assumed that ϵ is small in comparison with unity.

Since $\epsilon=\partial u/\partial x$, Eq. (1) can also be written in the form

$$\rho(\partial^2 u/\partial t^2)=T(\partial^2 u/\partial x^2), \tag{2}$$

where $T=d\sigma/d\epsilon$ is the modulus of deformation, elastic or plastic. The quantity T is considered to be a given function of the strain. The boundary conditions are $u=v_1 t$ for $x=0$, and $u=0$ for $x=-\infty$.

It is easily seen that a solution of the form

$$u=v_1[t+(x/c_1)] \tag{3}$$

with an arbitrary value of the velocity of propagation c_1, satisfies Eq. (2) and the boundary condition at $x=0$. For this solution the strain ϵ is constant and is equal to v_1/c_1.

A second solution is obtained by putting

$$T/\rho=x^2/t^2. \tag{4}$$

Since $T[=d\sigma/d\epsilon]$ is a given function of ϵ, Eq. (4) represents a solution for which ϵ is a function only of the variable $\xi=x/t$.

Assume $\epsilon=f(\xi)$; then the displacement u has the form

$$u=\int_{-\infty}^{x}\frac{\partial u}{\partial x}dx=\int_{-\infty}^{x}f(\xi)dx=t\int_{-\infty}^{\xi}f(\xi)d\xi \tag{5}$$

since $dx=td\xi$. By differentiation one readily obtains

$$\begin{aligned}\partial^2 u/\partial t^2&=(\xi^2/t)f'(\xi)\\ \partial^2 u/\partial x^2&=(1/t)f'(\xi)\end{aligned} \tag{6}$$

and substitution of Eqs. (6) in Eq. (2) shows that one of the two equations,

$$\rho\xi^2=T \tag{7}$$

or

$$f'(\xi)=0 \tag{8}$$

must hold. Equation (8) leads to the solution expressed by Eq. (3), whereas Eq. (7) gives the solution of Eq. (4).

The complete solution is obtained as follows:

(a) For $|x|<c_1 t$ the strain ϵ is constant and equal to ϵ_1;

(b) For $c_1 t<|x|<c_0 t$, where c_0 is the velocity of propagation of the elastic wave,

$$T(\epsilon)=\rho(x^2/t^2) \tag{9}$$

(c) For $|x|>c_0 t$, $\epsilon=0$.

The distribution of ϵ as a function of ξ is shown schematically in Fig. 1. The value of T for small values of ϵ—that is, within the elastic limit—is equal to E, Young's modulus of elasticity for the material. The elastic wave propagates with the velocity $c_0=(E/\rho)^{\frac{1}{2}}$. Between the plastic wave front, which is propagated with the velocity c_1, and the elastic wave front, the strain is variable, since every strain-increase from ϵ to $\epsilon+d\epsilon$ proceeds with a velocity equal to the specific value of $c=(T/\rho)^{\frac{1}{2}}$, corresponding to the strain ϵ.

The main problem is to determine the velocity c_1 of the plastic wave and the maximum strain ϵ_1 as a function of the velocity of impact v_1.

From Eq. (5) it follows:

$$v_1 = \frac{u(0,t)}{t} = \int_{-\infty}^{0} f(\xi)d\xi \tag{10}$$

and by changing variables

$$v_1 = -\int_0^{\epsilon_1} \xi d\epsilon$$

(See Fig. 4).

Thus, upon substituting for ξ from Eq. (7), Eq. (10) takes the form

$$v_1 = \int_0^{\epsilon_1} (T/\rho)^{\frac{1}{2}} d\epsilon. \tag{11}$$

Since T is a given function of ϵ, Eq. (11) determines ϵ_1 as a function of v_1.

If the deformation remains within the elastic limit, $T=E=$constant, and $v_1=\epsilon_1 c_0$, then $v_1=\epsilon_1(E/\rho)^{\frac{1}{2}}$. Hence, the stress σ is given by

$$\sigma_1 = E\epsilon_1 = \rho v_1 c_0. \tag{12}$$

Equation (12) is universally used for the calculation of the stress produced in an elastic body subjected to an impact velocity v_1. It appears that Eq. (11) replaces Eq. (12) in the case of a deformation beyond the elastic limit. If the stress remains within the elastic limit, there are two regions: for $|x|<c_0 t$, $\sigma=\rho v_1 c_0$ and for $|x|>c_0 t$, $\sigma=0$. In the case of plastic deformation, there are two fronts. Beyond the front of the elastic wave, $\sigma=0$; between the fronts of the elastic and plastic waves, σ increases gradually from $\sigma=0$ to a maximum value, $\sigma=\sigma_1$; and behind the front of the plastic wave, σ has the constant value σ_1, corresponding to a total strain ϵ_1 —elastic plus permanent—where ϵ_1 is given by Eq. (11).

For most materials $d\sigma/d\epsilon$ approaches zero for large values of ϵ, and at some particular value of ϵ, the material breaks. Hence the integral constituting the right-hand member of Eq. (11) has a maximum value, and one obtains a critical value of the velocity v_1. It can be expected that an impact with a velocity larger than v_1 will cause an instantaneous breakdown of the material.

III. EXTENSION OF THE THEORY; BAR OF FINITE LENGTH

The characteristic parameters that define the state of strain and motion of an element are the strain $\epsilon[=\partial u/\partial x]$, the stress σ, and the velocity of the element $v=\partial u/\partial t$. The equation of motion for the element is

$$\rho(\partial v/\partial t) = \partial\sigma/\partial x$$

or, using the definition of the velocity of propagation,

$$\partial v/\partial t = c^2(\partial\epsilon/\partial x). \tag{13}$$

On the other hand, from the relations $v=\partial u/\partial t$ and $\epsilon=\partial u/\partial x$ it follows that

$$\partial v/\partial x = \partial\epsilon/\partial t. \tag{14}$$

Equations (13) and (14) are equivalent to the differential Eq. (2).

The coordinate x of the element and the time at which the magnitude of the strain of this element is equal to ϵ and its velocity is equal to v can be considered as functions of ϵ and v. Then Eqs. (13) and (14) take the form

$$\begin{aligned} \partial x/\partial\epsilon &= c^2(\partial t/\partial v) \\ \partial t/\partial\epsilon &= \partial x/\partial v. \end{aligned} \tag{15}$$

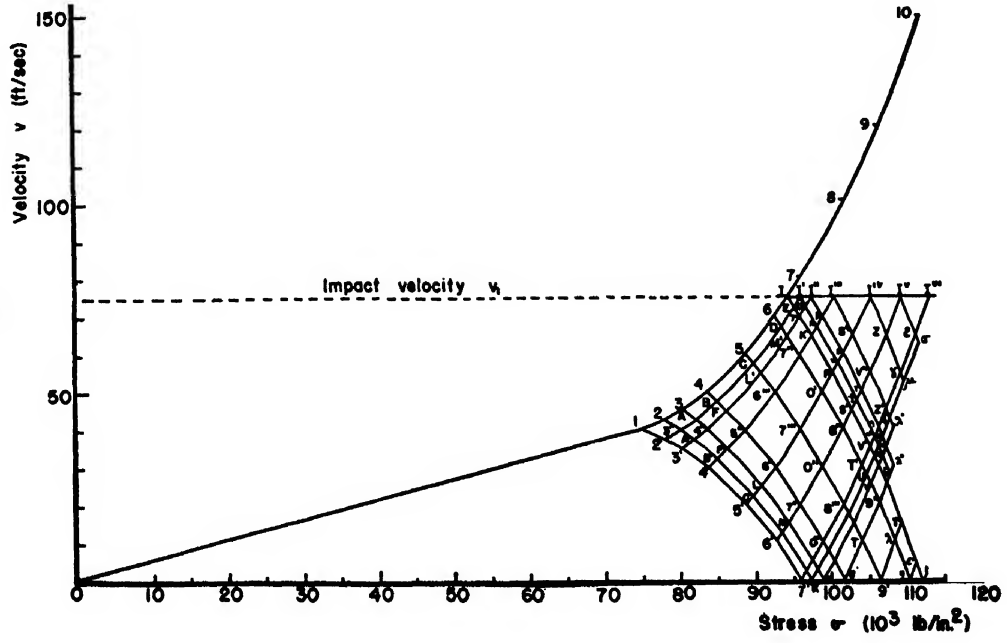


FIG. 2. Velocity *vs.* stress diagram for heat treated carbon steel.

Equations (15) become more symmetrical by the introduction of the function $\phi = \int_0^{\epsilon} c(\epsilon) d\epsilon$ and we obtain

$$\partial x/\partial \phi = c(\partial t/\partial v)$$
$$\partial x/\partial v = c(\partial t/\partial \phi). \tag{16}$$

The process of propagation can be represented in two planes. The first, with x and t as coordinates, is the Lagrangean or physical plane; the second plane is the velocity plane, in which ϕ and v are used as coordinates. The system of Eqs. (16) has fixed characteristics in the $v\phi$ plane. Hence, a relatively simple graphical method can be developed for the solution of a broad class of impact problems in the plastic range; for example, the case of a bar of finite length with one fixed end and hit at the free end. In Figs. 2 and 3 are typical (σ, v) and (x, t) diagrams for a specimen of heat treated carbon steel subjected to an impact velocity of 75 ft./sec.

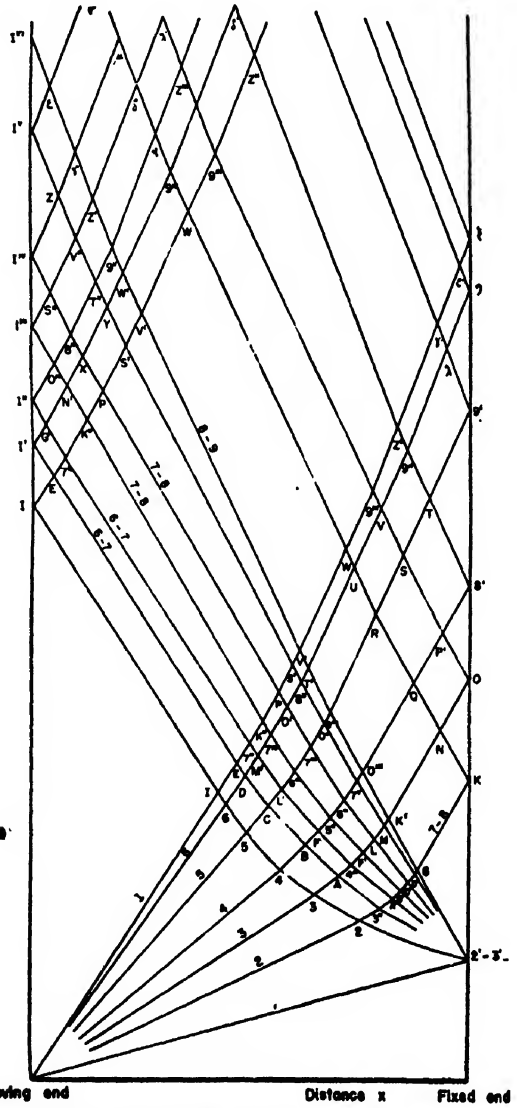


FIG. 3. Lagrange diagram for specimen of heat treated carbon steel. Impact velocity 75 ft./sec.

IV. EXPERIMENTS ON THE PROPAGATION OF PLASTIC STRAINS IN A LONG SPECIMEN SUBJECTED TO TENSION IMPACT

1. Experimental Technique

The first experiments on the propagation of plastic strain in tension were made in view of verifying the three following points resulting from the theory: (1) the existence of a plastic wave front of given amplitude; (2) the relation between the amplitude ϵ_1 of the plastic wave front and the velocity of impact v_1 (Eq. (11)); (3) the distribution of the plastic strain between the plastic and the elastic front as given by Eq. (9). The specimens used in these experiments were annealed copper wire about 100 inches long and 0.071 inch in diameter. The specimens were mounted vertically in a special testing machine built by Dr. D. S. Clark at the California Institute of Technology. The impact was produced by a hammer guided between two rails and accelerated by prestretched rubber bands. The maximum velocity attainable was approximately 200 ft./sec. The velocity was measured by a suitable electric device and a cathode-ray oscillograph. On every specimen equidistant marks were made with one inch spacings. After the test, the plastic strain was determined by measuring the displacement of each mark. Experiments were also performed in compression impact. In this case, the specimens were cylinders 12 inches long and about $\frac{3}{8}$ inch in diameter.

To control the duration of impact, the following device was employed: the bottom end of the wire is attached to a rigid piece A, as shown in Fig. 4. A vertical rod B, resting on the bottom frame of the machine, fits loosely into the tubular part of A. When a hammer H hits the piece A, the specimen elongates until A reaches the rod B. The piece A contains a circular notch N, and the rim of A breaks off at this notch after A has traveled the distance D and comes to rest on the rod B. The purpose of this arrangement is to allow the hammer to continue to move downward and also to dissipate some of its remaining energy. However, no kinetic energy is transferred to the specimen after A reaches B. The time of impact is therefore the distance D divided by the velocity of the hammer during the process of elongation.

The static stress-strain curve of one of the copper wires tested in impact was used to compute the values of $d\sigma/d\epsilon$ as a function of ϵ and the values of the velocity of propagation c, as given by the formula $c = [1/c(d\sigma/d\epsilon)]^{\frac{1}{2}}$, were then computed and a curve of c *versus* ϵ was obtained (Figs. 5 and 6). From this curve, the value of the quantity v_1 (the velocity of impact corresponding to a plastic front of amplitude ϵ_1) was calculated for each value of ϵ_1 (Fig. 7). The largest value of v_1 is obtained for the largest possible value of (16 percent in this case) and is equal to 150 ft./sec. The impact velocity of 150 ft./sec. is therefore the "critical velocity" for this material. An impact with a higher velocity

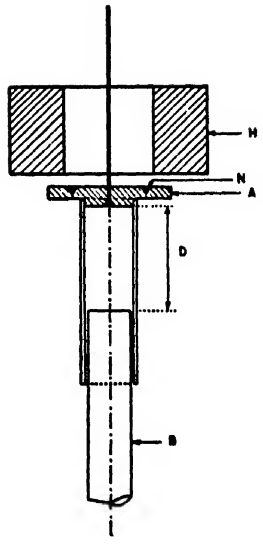


FIG. 4. Experimental device used to stop the impact after a given deformation of the specimen has been reached.

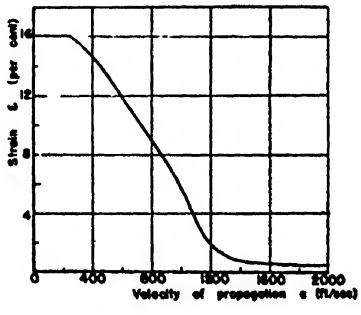


FIG. 5. Strain ϵ *versus* velocity of propagation c as calculated from the static stress-strain curve

must produce an instantaneous breakdown of the specimen.

As mentioned above, marks were made on the specimen at intervals of one inch. The distance between the origin and each mark was measured before and after the test. If x_n denotes the distance from the origin to a mark n before the test, and if x_n' is the distance between the two points after the test, the difference $x_n'-x_n$ represents the total elongation of the specimen between these two points. The values of $x_n'-x_n$ were then plotted *versus* x. The slope at any point of the curve so obtained gave the value of the permanent strain at that point. This way of measuring the strain has been found to be the most practical one. It was not necessary to make the distinction between permanent strain and actual strain, because the elastic recovery was relatively too small to be taken into consideration.

2. Results of the High Velocity Impact Tests

The results are given in three separate subsections—(i), (ii), and (iii)—in order to show to what extent they agree with the three principal theoretical results which were expressed in Section II.

(*i*) *Existence of a Plastic Wave Front*

The first series of tests was made to establish that the amplitude of the plastic front is a function of the velocity of impact alone. This amplitude remains constant while the elastic front and the plastic front travel along the specimen. In these particular experiments, the velocity of impact was always 92.50 ft./sec., but the durations of the impact varied. To control the duration of impact the two pieces A and B, shown in Fig. 4, were placed at a proper distance apart. The curves in Fig. 8 give the distribution of the strain along the specimen. They indicate clearly that a plastic front of a given amplitude ϵ_1 is revealed by the experiments, as predicted from the theory.

(*ii*) *Relation Between the Velocity of Impact and the Amplitude of the Plastic Wave Front*

In this series of tests, the velocity of impact was varied from one test to the other. The total elongation was not necessarily the same for all the tests. The stopping device was adjusted in such a manner that during the impact the plastic front traveled a distance of between 20 and 40 inches. Figure 9 gives the distribution of the strain along the specimen in each case. The experimental values of ϵ_1 corresponding to each velocity tested are listed in Table I. For comparison with the theory, the experimental values are plotted as points in Fig. 7, whereas the solid curve represents the result of the theoretical computation. The agreement between the experimental results and the predetermined curve is fairly good. The highest velocity of impact used was 171 ft./sec. For that velocity, the specimen broke within the first inch, indicating that a velocity of 171 ft./sec. is above the critical velocity, as predicted by the theory. It must be pointed out that a considerable reduction of area was observed, which indicates that the rupture was not brittle. In addition, a plastic strain of relatively small intensity (less than

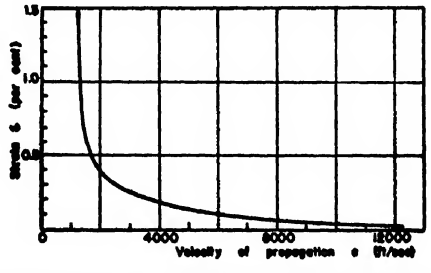


FIG. 6. Continuation of the curve of Fig. 5 up to the elastic point.

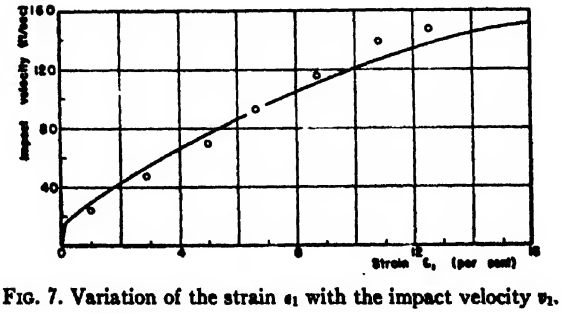


FIG. 7. Variation of the strain ϵ_1 with the impact velocity v_1.

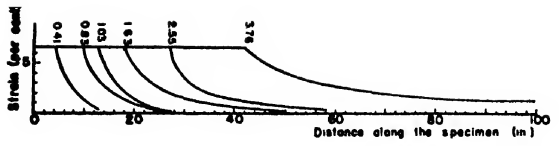


FIG. 8. Distribution of the strain ϵ along the specimen. Velocity of impact 92.5 ft./sec. The duration of impact is indicated on each curve in msec.

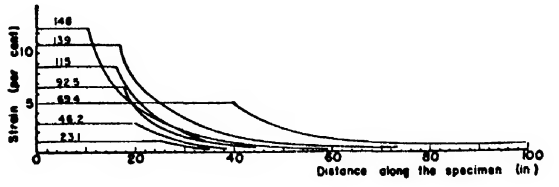


FIG. 9. Strain distribution curves for annealed copper specimens subjected to different impact velocities, indicated on each curve in ft./sec.

two or three percent) was measured along the wire over a distance of about 20 inches from the point of rupture. It is probable that a plastic wave started to propagate along the specimen during the time required to produce rupture.

(iii) *Shape of the Plastic Wave and Velocity of Propagation of the Plastic Front*

The distribution of the plastic strain along the wire between the plastic and elastic fronts may be calculated from Eq. (9), using the curves of Figs. 5 and 6. The dashed curves in Fig. 10 represent theoretically computed amplitude distributions of a plastic wave after a time interval equal to 0.83 msec.; the assumed velocity of impact, v_1, is 92.5 ft./sec. and ϵ_1 is 6.6 percent. The solid curve in Fig. 10 represents plots of the measured distribution of the strain. It is seen that the experimental and theoretical curves deviate considerably. This deviation is due to some inaccuracy in the value of the duration of impact, and especially to the perturbating effects brought by the sudden stopping of the impact. When the hammer which pulls the end of the wire with a constant velocity is stopped, the portion of the specimen that is in motion is not stopped instantaneously. During the deceleration period, the kinetic energy stored in the specimen is transformed into strain energy, and the total extension is larger than the distance D through which the hammer has been in contact with the specimen.

TABLE I. Results of impact tests on 80 inches long annealed copper specimens.

Impact velocity (ft./sec.)	Uniform strain (percent)	Duration of impact (msec.)	Velocity of propag. of uniform strain (ft./sec.)
23.1	1.0	1.70	1230
46.2	2.9	1.47	1130
69.4	5.0	3.37	1000
92.5	6.6	1.63	910
115.0	8.7	1.55	860
139.0	10.8	1.94	730
148.0	12.5	1.40	630
171.0		Rupture	

The measured shape of the plastic wave given in Fig. 10 is greatly influenced by this "stopping effect" and cannot logically be compared with the theoretical curve that represents the shape of the wave at a certain instant during impact. The stopping effect was analyzed theoretically by the authors and especially by Drs. F. Bohnenblust and J. Charyk, who collaborated in the work reported in this paper. These investigations will be published separately. They materially reduced the variance between computed and measured values.

The reflection of plastic waves at the fixed end of a specimen was also investigated theoretically and experimentally. A typical result of this investigation is shown in Fig. 11, in which a measured strain distribution curve obtained in a compression impact test is compared with a computed curve. The two curves are not in complete agreement, but the character of the observed strain distribution curve is accounted for by the theory. It may also be pointed out that the deviation between the two curves certainly reflects some influence of the strain rate, showing that the assumption of a stress-strain curve independent of the rate of strain is not entirely justified.

A very noticeable effect of the rate of strain on the elastic limit has been observed for iron specimens subjected to both tension and compression impact. The most striking result of impact tests on iron specimens

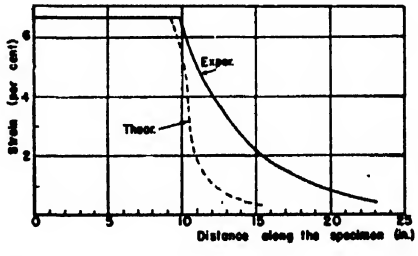


FIG. 10. Theoretical and experimental strain distribution curves for annealed copper specimen. Impact velocity 92.5 ft./sec. Duration of impact 0.83 msec.

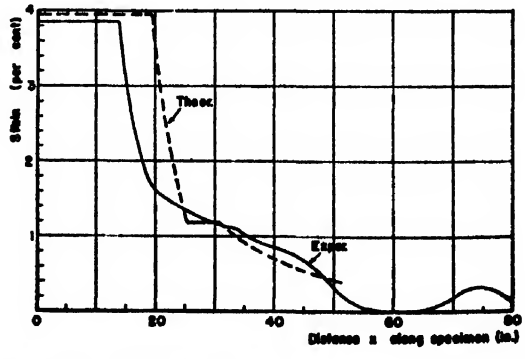


FIG. 11. Experimental and theoretical strain distribution curves for a copper wire 0.08 inch in diameter. Impact velocity 62.5 ft./sec., duration of impact 1.3 msec.

was that no plastic strain occurred near the moving end until the stress wave reached a value of the order of three times the static yield stress. The measured strain distribution curves also showed a different character, as shown in Figs. 12 and 13. It can be observed that in the case of compression the permanent set is concentrated in the neighborhood of the ends of the specimen and the center is free from plastic strain. In the tension tests there were intervals of permanent sets alternating with undisturbed regions. This is probably due to irregularities in the specimens whose influence was

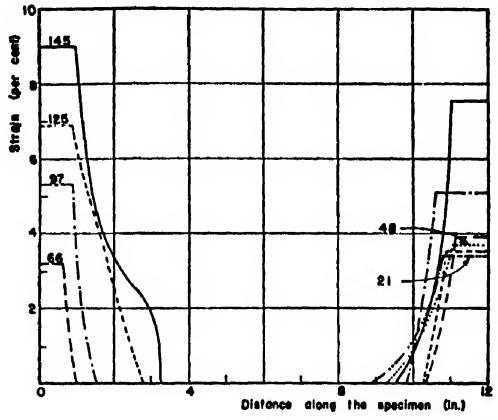


FIG. 12. Strain distribution curves for annealed mild steel specimens subjected to compression with different impact velocities, indicated on each curve in ft./sec.

enhanced by the unstable character of the tension process.

3. Force-Time Relation at the Fixed End of a Specimen Subjected to Tension Impact

An important contribution of the theory of plastic wave propagation presented in this paper is a clear explanation of the dependence on time of the force measured at the fixed end of a specimen subjected to tension impact. Force *versus* time diagrams recorded at the fixed end of the specimen have been interpreted by previous investigators in the field of high velocity impact testing as dynamic stress-strain curves. The transformation of the time axis into a strain axis infers that the strain is uniform all along the specimen at any instant during impact. The theory of strain propagation shows that for impact velocities greater than approximately 10 ft./sec., this hypothesis is not justified, and a stress-strain relation cannot be obtained in such a simple way. In order to show this, the force acting at the fixed end of a specimen subjected to tension impact was computed as a function of time by the theory outlined in Section III. This force was also measured in a series of experiments made with different types of specimens. The force was recorded by means of a resistance sensitive strain gauge in conjunction with a cathode ray oscillograph. Two records obtained for copper specimens tested with an impact velocity of 100 ft./sec. are reproduced in Fig. 14, together with a theoretical stress *versus* time diagram. The agreement between theoretical and experimental curves is very satisfactory as far as their shape is concerned. The experimental curve, however, seems to be shifted to the left. This discrepancy could be explained by assuming that any strain has a velocity of propagation slightly higher than that based on the static stress-strain curve. This gives further evidence of the influence of the rate of strain on the relation between stress and strain. In any case, the interpretation of such curves as those of Fig. 14 as dynamic stress-strain curves is completely fictitious.

4. Plastic Strain Propagation and High Velocity Impact Tests

High velocity tensile impact tests have been reported by several investigators.[1–6] Some of the results[1] have shown that the energy absorbed before rupture decreased when a certain impact velocity was reached.

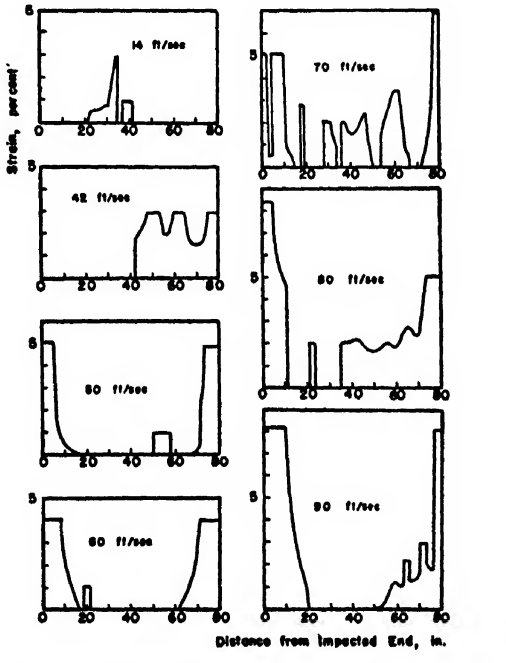


FIG. 13. Strain distribution for annealed iron wire in tension for different impact velocities.

However, the observed drop in the amount of absorbed word was not accompanied by a decrease in elongation

[1] H. C. Mann, Proc. A.S.T.M. 36, 85 (1936).
[2] D. S. Clark and G. Datwyler, Proc. A.S.T.M. 38, 98 (1938).
[3] D. S. Clark, *The Influence of Impact Velocity on the Tensile Characteristics of Some Aircraft Metals and Alloys*, NACA Technical Note No. 868.
[4] M. Manjoine and A. Nadai, Proc. A.S.T.M. 40, 822 (1940). Trans. A.S.M.E., J. App. Mech. 8, A-77 (June, 1941).
[5] DeForest, MacGregor, and Anderson, Metals Technol. 8, 1 (December, 1941).
[6] E. R. Parker and C. Ferguson, Trans. A.S.M. XXX, No. 1, 68 (March, 1942).

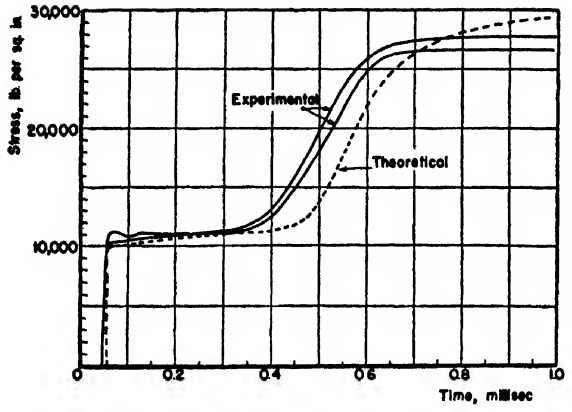


FIG. 14. Experimental and theoretical stress-time curves in tension for fixed end of annealed copper. Gage length eight inches; impact velocity 100 ft./sec.

and no explanation was offered for this phenomenon. Other investigators did not observe any drop in energy absorption, even with impact velocities as high as 200 ft./sec.

After the discovery of the laws of propagation of plastic deformation, it became evident to the authors that both the elongation and the energy should decrease more or less suddenly when the critical velocity is reached. The first tests made with long copper wires and described in Section III clearly established the existence of a critical velocity, as predicted by the theory. The rupture of the specimen, however, at velocities above the critical, was still of the ductile type, and some appreciable strain due to necking could still be measured near the end of the specimen subjected to impact. The elongation due to necking and the energy absorbed by the necking process are relatively important when the test specimen is short, or rather when the ratio between the length and the diameter of the specimen is small. Systematic experiments indicated that, in order to observe a definite drop in both energy and total elongation beyond the critical velocity, specimens should have a ratio of length to diameter at least equal to 20. These results furnished a clear explanation why a critical velocity was not observed by previous investigators, who made high velocity impact tests on short specimens.

One of the fundamental questions in high velocity impact tests is the influence of the rate of strain. The investigations of the authors clearly showed that such an influence exists. In fact, whereas the general character of the impact process was fairly well described by a theory neglecting the influence of the rate of strain, there was always a certain systematic discrepancy pointing to such an influence. However, it should be emphasized that in high speed impact tests the rate of strain varies from one point to the other and is a function of time. Computations resulting from the theory of wave propagation have shown that the average rate of strain calculated on the basis of uniform strain is far from the actual rate of strain, even when the velocity of impact is as low as 10 ft./sec. It is therefore not logical to utilize tension impact tests to study the influence of the rate of strain on the properties of metals.

Finally, it appears that for materials having a yield point like soft iron, the interpretation of high velocity impact tests is made rather difficult because of the phenomena reported in Section IV-2. The authors believe that the understanding of these phenomena requires a deeper insight into the whole question of plastic deformation and the physics of solids.

V. PRACTICAL APPLICATIONS

As previously stated, the theory of propagation of plastic deformation was developed for the understanding of the resistance of structures against impact. Hence the first application which appears to be important is the development of rules which enable the designer to compute the amount of energy which can be absorbed by a given structure. It has been shown quite clearly for the first time that the amount of energy is a function of the impact velocity, with a rather sudden decrease of the energy absorption beyond a critical velocity. The theory has already been extended to beams subjected to transverse impact. A class of ballistic problems which have not been sufficiently discussed from the point of view of plastic strain should be mentioned. Certain aspects of the problem of penetration of projectiles through metal plates are connected with plastic wave propagation. It seems, for example, that the notion of critical velocity could lead to the discrimination of the cases in which the penetrating projectile merely pushes aside the material from those in which a plug is separated from the back of the plate. Finally, the notion of plastic strain propagation might bring about a better understanding of some anomalies encountered in high speed machining.

A Kerr Cell Camera and Flash Illumination Unit for Ballistic Photography

H. F. Quinn, W. B. McKay, and O. J. Bourque
Canadian Armament Research and Development Establishment, Valcartier, Quebec, Canada
(Received March 30, 1950)

This paper describes an experimental model of a combination Kerr cell camera and flash illumination unit designed specifically for the photography of projectiles under conditions of strong ambient illumination. Specific cases wherein such conditions are found are considered and the applicability of the Kerr cell as a secondary protecting shutter is described. A description is given of a pulse forming network—pulse transformer circuit designed to apply an operating voltage pulse to a Kerr cell of thirty-six (36) kilovolts amplitude and one point nine (1.9) microseconds duration. Limitations inherent in the present equipment are considered together with proposals for overcoming these defects in future models. Typical results are presented, including (a) a picture taken with the unit showing a twenty-millimeter shot at the time of emergence from the muzzle of a gun, and (b) a picture of a thirty-seven millimeter shot taken shortly after penetration of a metal target plate.

THE "single shot" photography of projectiles in flight by means of a high intensity flash lamp of the Egerton type is now a well-established technique in most ballistic research laboratories. In this technique, the exposure time of each picture is determined by the effective photographic duration of the particular flash lamp employed and is normally of the order of a microsecond. Such an effective exposure is entirely adequate to "stop" all but the highest speed projectiles in flight. It is a conventional procedure to use a good camera of the Eastman "Speed Graphic" type with a shutter setting of 1/100 second, the shutter opening being suitably controlled by electrical circuits of various degrees of complexity to ensure that the shutter is open for some short interval before, while, and for a short interval after the projectile is in the field of view of the camera. Auxiliary electrical circuits are provided whereby the projectile, in passing some predetermined position in the field of view, triggers the flash lamp and the required picture is taken. As it is a common procedure to employ very fast film in these applications, some means must be provided to protect the film from fogging caused by ambient daylight even with the short exposure of 1/100 second and, in many cases, both the camera and the light source are positioned in a darkened hut through which the projectile passes via suitably placed apertures in the walls. A typical photograph of a conventional shell taken under such conditions is shown in Fig. 1.

Frequently, however, it is required to obtain pictures of projectiles under conditions where the subject is either completely or partially surrounded by a glowing medium of considerable luminosity. Examples of such cases are:

(a) A projectile emerging from the muzzle of a gun and surrounded by high intensity luminous gases (muzzle flash).

(b) A projectile striking an armor plate and surrounded by an atmosphere of glowing, incandescent particles of metal which have been stripped from the plate by the projectile impact.

A particularly good example of the ambient flash illumination produced in (b) is illustrated in Fig. 2, a photograph taken at this laboratory of an aluminum-capped projectile attacking an armor plate. It should be observed that the extraneous illumination present in such cases would not only tend to screen the projectile from the external photographic flash but also, by its own inherently high luminosity and relatively long duration (of the order of 1/100 second) would very completely fog the film during the 1/100 second opening of the mechanical shutter. Such total fogging will be observed in Fig. 2 in the regions of maximum interest to the terminal ballistician namely at the point of impact (A) on the front surface of the plate and at the point of emergence (B) at the rear. An attempt to employ a faster mechanical shutter would be entirely ineffective: the fastest mechanical shutter is one of the focal plane type, models of which have been built with quoted shutter speeds of 1/5000 second. Not only would such a shutter prove almost impossible to synchronize with the external flash source but would also fail to protect the film from fogging due to the ambient illumination from regions A and B even with the above short exposure. It is, then, necessary to resort to a relatively inertia-less, non-mechanical shutter. Some wartime investigations

Fig. 1. Conventional "Microflash" photograph of a shell in flight.

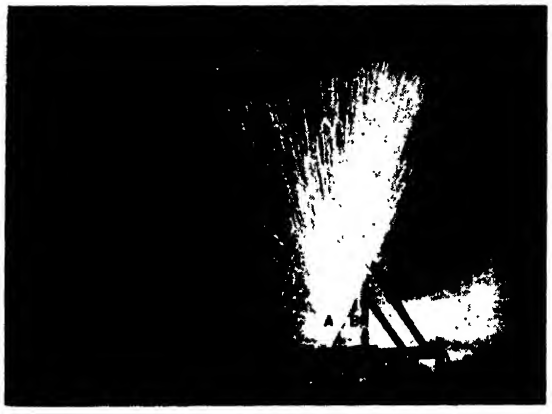


FIG. 2. Ambient flash illumination produced by an aluminum-capped projectile striking an armor plate.

performed in Germany and also in Great Britain suggested the possibility of utilizing the practically inertialess Kerr Cell electro-optical shutter in series with a conventional mechanical shutter to exclude the effect of the ambient flash. Security limitations do not, unfortunately, permit reference either to these investigations or to the results obtained therein. It would, of course, be imperative to effect a precise synchronization between the opening of the Kerr shutter and the initiation of the external flash lamp to insure that the illumination caused by the latter was fully utilized. At the end of the external flash, the Kerr shutter would close to exclude the ambient flash and, finally, the mechanical shutter would close to eliminate daylight fogging of the film noting particularly that the crossed polaroid sheets commonly used in most present day Kerr cells are reasonably transparent in the red portion of the spectrum. We desire to report on a preliminary, experimental model of a simple, self-contained unit incorporating both a Kerr cell camera and its associated flash illumination source for use in such applications as have been described above. The use of an artificial transmission line type of pulse-forming network together with a pulse transformer to obtain a substantially rectangular voltage pulse of two microsecond duration and 36 kilovolt peak amplitude for operating the Kerr cell is believed to be a new feature. It is further desired to point out some major disadvantages in the present equipment and to indicate how these may be overcome in a future model.

THEORY OF THE KERR CELL SHUTTER AND PREVIOUS APPLICATIONS

The theory and method of operation of the Kerr cell electro-optical shutter has been extensively discussed in the physics literature of the past fifty years; probably the most modern and satisfactory treatment being that due to Zarem.[1] It is not, therefore, proposed to include this information in the present report. It is, however, of interest to mention some of the uses to which the Kerr shutter has been applied by previous investigators and to consider the various aspects in which the present usage of the cell differs from them. Accordingly, Table I has been constructed in which several of the former applications have been listed together with the duration of the effective photographic exposure obtained in each case.

It is to be noted that, in all but the present application, the subjects to be photographed themselves provided all the necessary illumination to enable reasonably good pictures to be obtained notwithstanding the high light absorption losses in the active fluid (usually nitrobenzene) of the cell. The present usage of the Kerr shutter differs markedly from the others in that the useful light must be provided from an external, high intensity flash source, the light flash being accurately

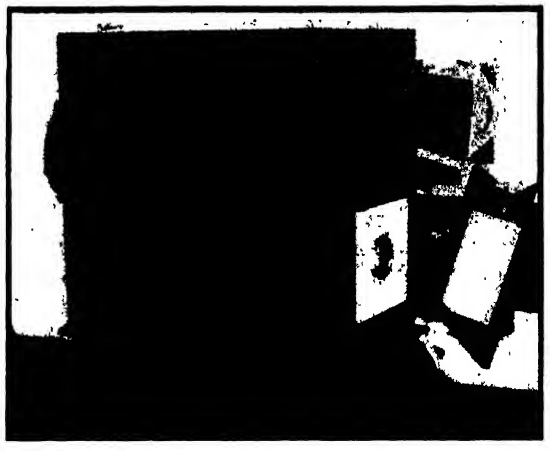


FIG. 3. Kerr cell camera and flash illumination unit.

synchronized with the opening of the shutter. Further, in the present application, the effective opening of the shutter need not be set precisely equal to the desired exposure time, provided it is at least sufficient to admit the external flash illumination and it is the effective photographic duration of the light flash that determines the exposure and not the opening of the Kerr shutter. Finally, the present application necessitates the use of a Kerr cell of much greater aperture and angular field of view than those employed by the above investigators; this, in turn, requires the use of much higher driving voltages in order to obtain a satisfactory opening of the shutter.

DESCRIPTION OF THE APPARATUS

A. Optical Arrangements

A photograph of the complete unit is shown in Fig. 3. The mirrors mounted on the front of the unit enable the latter to be suitably protected from flying fragments in the case of armor plate investigations and from the effect of muzzle blast in applications where it is desired to photograph projectiles emerging from a gun muzzle.

[1] Zarem, Marshall, and Poole, Elect. Eng. 68, No. 4, 282 (1949).

The general arrangement of equipment as used in the latter application is shown in Fig. 4 wherein the unit is seen in position beside the gun: the lower mirror enables the substantially parallel beam of light from the flash lamp to be trained on the muzzle while the upper mirror allows the camera to view the projectile on emergence. Initiation of both the Kerr camera and the flash lamp is provided by electrical contact between the emerging projectile and a stout wire tightly strung some one-half inch clear of the muzzle.

The camera proper consists of a "Baco" mounting and bellows together with a suitable adapter which permits the mounting of a $f/2.5$ "Aero Ektar" lens having a focal length of 8 inches. In this preliminary, experimental model the mechanical shutter was dispensed with and all tests were carried out at night inside a darkened gun shed. Suitable synchronization of a conventional, mechanical shutter from the gun itself would, however, present no great problem such an arrangement having been frequently employed by this laboratory in the past.

FIG. 4. Experimental arrangement for photography of a projectile at point of emergence from a gun muzzle.

Insofar as the dimensions of the electrodes are concerned, the Kerr cell follows exactly a German wartime model specified to operate at a peak voltage pulse of 36 kilovolts. As it was anticipated that considerable investigation might have to be carried out before the optimum design of the cell was finalized, the use of the more convenient (and much more expensive) all-glass cell was avoided and, to date, all the experimental cells have been constructed from one-inch Bakelite sheet and ordinary quarter-inch plate glass. All joints have been carefully sealed with ceresin wax and care has been taken to remove the nitrobenzene from the cell after each trial to eliminate the small but noticeable solution of this wax in the fluid. Type J Polaroid film pressed between the cell windows and external glass plates has been used to date for the linear polarizers but it is intended shortly to replace this by Type H film laminated in glass. The particular Kerr cell used in the first successful trial of the unit is shown in Fig. 5. The cell is placed immediately in front of the camera lens in the unit. It is not possible at the present time to calculate

TABLE I.

Author	Application	Effective exposure microseconds
Beams[a]	Study of the successive appearance of spectral lines in spark discharges.	0.001 to 0.1
Dunnington[b]	Photographic study of the early stages of the spark discharge.	0.002
Anderson[c]	Measurement of the velocity of light.	—*
Froome[d]	Fundamental studies of cathode spot formation in the arc discharge.	0.1
Zarem[e]	Fundamental study of the process of exploding wires.	0.04
Present authors	Ballistic photography.	2.00

* In this application, the Kerr cell was utilized as a light modulator valve.
[a] J. W. Beams, J. Opt. Soc. Am. 5, 597 (1926).
[b] F. G. Dunnington, Phys. Rev. 38, 1506 (1931).
[c] W. C. Anderson, Rev. Sci. Inst. 8, 239 (1937).
[d] D. K. Froome, J. Sci. Inst. 25, No. 11, 371 (1948).
[e] See reference 1.

accurately the theoretical maximum transmission of the Kerr cell when used with the present light source, the spectral distribution of energy in the flash being unknown. It is, however, planned to photograph the spectrum of the discharge lamp using Super XX film (the film actually used in applications of the equipment) and to substitute in the transmission formula the appropriate value of the Kerr constant corresponding to the spectral region of maximum light intensity. Some idea of the performance of this particular cell can, however, be obtained by substitution in the transmission equation of values for both the Kerr constant and the

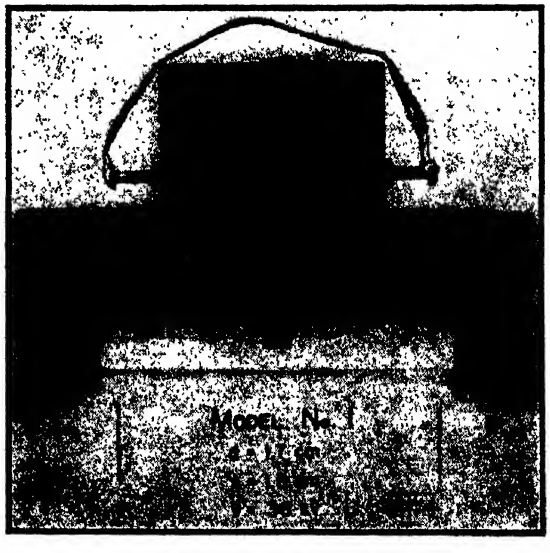


FIG. 5. First experimental Kerr cell.

absorption coefficient for a wave-length reasonably well in the center of the optical spectrum: let the following numerical values be used:

Kerr constant for nitrobenzene at λ
$= 5460$ angstroms $= 410 \times 10^{-7}$.

Absorption coefficient at $\lambda = 5460$ angstroms $= 1/15$.

The transmission equation, corrected for absorption in the cell, can be given in the form:

$$\text{Percent transmission, } T = 50 \exp -\rho \times \sin^2\left(\frac{\pi \times BV^2}{d^2}\right), \quad (1)$$

where x (cm) is the axial length of the cell plates, B the Kerr constant, d (cm) the plate separation, ρ the absorption coefficient and V (statvolts) the applied voltage. Substituting the above numerical values and the cell dimensions we obtain the result, $T = 37.1$ percent, the losses inherent in the linear polarizers being neglected.

In the present apparatus, the components available permitted operation of the cell only up to the originally specified limit of 36 kilovolts. Provided the dimensions x and d be fixed, it is apparent from Eq. (1) that a maximum possible percent transmission of 44.3 percent would be obtained for $V = 41.9$ kilovolts.

B. Electrical Circuits

The complete electrical circuit for the Kerr camera and its associated flash lamp is given in Fig. 6. Referring to Fig. 6, it will be seen that, when the high voltage supply is energized, one side of the main spark gap S charges to $+9$ and the other to -9 kilovolts. These potentials also appear across the flash energy storage capacitors C-4, C-5 and the pulse-forming network, PFN. On shorting the grid of the mercury thyratron V-3 to ground, a negative pulse of 4.5 kilovolts amplitude appears at the center electrode this being sufficient to trigger the main gap. Proper initiation of the main spark was found to be ensured by adding the capacitor C-3 across the gap. As soon as the gap fires, a total potential of 18 kilovolts appears across the flash lamp, a General Electric type FT-125, and this breaks down immediately, simultaneously discharging the pulse-forming network through the primary of the pulse-transformer T-5 and also the storage capacitors through the lamp itself. This arrangement of the circuit ensures a consistent synchronization between the camera and the flash lamp. A suitable purely resistive load of 640 ohms is connected in parallel with the Kerr cell across the

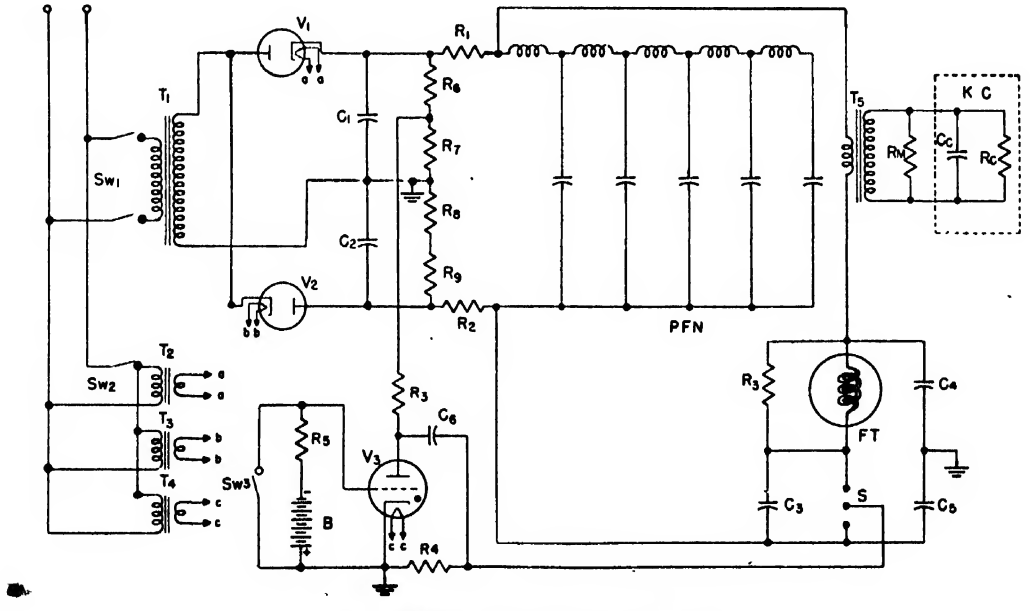


FIG. 6. General circuit diagram of the unit.

Referring letters	Components	Referring letters	Components
B	4–45 volts dry cell in series.	R_4	20 M at 20 kv.
C_1, C_2	Storage Capacitors, 0.25 μf at 25 kv.	R_5	10 K.
C_3	Capacitor, 0.0011 μf at 20 kv, mica.	R_6, R_7	10 M at 10 kv.
C_4, C_5	Flash energy storage capacitors, 1 μf at 10 kv.	R_8, R_9	10 M at 10 kv.
C_6	Capacitor, 0.0011 μf at 20 kv, mica.	R_M	Matching resistor—640 ohms.
FT	Flash tube, G.E. Type FT-125.	S	Control spark gap.
$K.C.$	Equivalent electrical circuit of Kerr cell.	Sw_2	Initiating contact switch.
PFN	Pulse-forming network—Zn = 40 ohms, $T = 1.9$ microsec.	T_1	Transformer—115, 7000 v.a.c.; 20 ma.
Cc	Capacitance of Kerr cell, $\simeq 55$ $\mu\mu$f.	T_2, T_3, T_4	Transformer, filament.
Rc	Resistance of Kerr cell, $\simeq 60$ K.	T_5	Transformer, pulse, step-up 1:4.
R_1, R_2	10 M at 10 kv.	V_1, V_2	High vacuum rectifier, type: 8013.
R_3	20 M at 20 kv.	V_3	Mercury thyratron, British CV-22.

secondary of the pulse transformer to match the transformer primary to the characteristic impedance of the pulse-forming network (40 ohms, in this particular case). While it will be noted that the voltage pulse at the transformer secondary cannot begin to rise until the initiation of the flash lamp, and that, by this means an approximate synchronism of the cell opening and the light flash is achieved, it is immediately apparent that the "rise time" to peak of the secondary voltage of a pulse transformer is quite considerable particularly when the latter works into an essentially capacitive load such as the Kerr cell (capacity of the order of 55 micro-microfarads). Lack of a suitable high speed single transient type oscilloscope has prevented the determination of the "rise time" in the present transformer. The photographic evidence is, however, that the "rise time" is, unfortunately, long enough to limit the useful light to that from the tail-end of the flash. There is no apparent simple means of eliminating this objectionable feature in the present equipment but, in future models, it is proposed to provide separate triggering of the cell and the lamp together with a precision time delay in between the two trigger pulses whereby the secondary voltage of the pulse-transformer will be given adequate time to rise to peak before the initiation of the flash. The fact that the photographic results obtained to date with this relatively crude apparatus have been so promising would seem to augur well for the success of the new equipment which is presently under construction.

It is important to note that the pulse transformer application is particularly appropriate to the present problem wherein a cell opening of one or two microseconds duration is the shortest required. The production of a pulse transformer with a four or five to one step-up ratio capable of providing a shorter voltage pulse of, for example, 0.1 microsecond to a capacitive load of some 60 micro-microfarads in parallel with the normal stray winding capacity (itself normally of the order of 40 mmf) would be a practical impossibility. Since, however, recent pulse transformers (with step-up ratios as great as one to seventeen) have been designed to handle two or three microsecond pulses, the use of a suitable pulse transformer should enable the operation of Kerr cells of much larger aperture with a consequent improvement in the photographic properties of the system as a whole. Furthermore, by the use of larger aperture Kerr cells, it should be possible to effect a substantial reduction in the capacity of the cells with a consequent reduction in the required power rating of the pulse transformer.

Concerning the several components used in the circuit, the pulse-forming network is of conventional design having been salvaged from a wartime radar set: it is rated at 25 kilovolts, has a specified pulse-width of 1.9 microseconds and a characteristic impedance of 40 ohms. The energy storage capacitors C-4 and C-5 were chosen from the standpoint of availability and insulation rating rather than for their particular suitability for flash duty; in fact they are high voltage filter capacitors. These capacitors will, of course, be replaced in future models.

Considerable difficulty was encountered throughout our investigations in the search for a suitable flash tube which would absorb the necessary power without disintegrating and which would also trigger from a suddenly applied surge voltage. While the FT-125 lamps

FIG. 7. Kerr cell photograph of a 20-mm shell emerging from gun muzzle.

finally selected are required to operate at a power level greatly in excess of their normal commercial rating, these lamps have given excellent results to date and, after some hundred flashes, still show no apparent signs of deterioration.

The pulse transformer was constructed in this laboratory: as it was initially desired merely to examine the applicability of the electrical circuit, no great effort was made in the construction of the device to minimize either the leakage inductance or the winding capacitance. A silicon steel core obtained from a surplus radar pulse transformer was utilized. The primary winding consists simply of two 40 turn windings each wound on one leg of the core and connected in parallel, the secondary consisting of two 80 turn windings each wound over one primary section and connected in series to give a voltage transformation ratio of 1 to 4. The transformer is immersed in oil in an old automotive battery case. Since the first promising results were obtained with the present equipment, orders have been placed with a manufacturer for a suitable pulse-forming

network having a characteristic impedance of 200 ohms and a pulse-width of 2 microseconds together with an associated pulse transformer having a 1 to 5 step-up ratio. This combination should enable a cell to be pulsed with a peak voltage up to the limit of 50 kilovolts.

EXPERIMENTAL PROCEDURE AND INITIAL RESULTS

The first significant and promising results were obtained in an application of the equipment wherein it was desired to obtain pictures of a twenty (20)-millimeter cannon shot just emerging from the muzzle of the gun for the purpose of determining the amount of propellant gas leakage past the driving band of the projectile. The experimental arrangement has already been illustrated in Fig. 4. The filament and high voltage power supplies were switched on and the thyratron grid connected to the contact wire at the muzzle; since the gun itself was grounded, the emerging projectile connected the grid to ground on contact with the wire. When these preliminary arrangements had been completed, the gunshed was darkened and the protecting slide removed from the film holder. Immediately after firing the gun, the slide was replaced in order to protect the film from stray illumination. A typical photograph of the emerging projectile (one of a series of twenty pictures obtained) is shown in Fig. (7) while Fig. (8) shows the opaque gas cloud surrounding the projectile when the latter had just cleared the muzzle. It should be remarked that these and all the pictures obtained with the Kerr camera have been recorded on Super-Pan Press, Type B film; the film records have in all cases been processed for nine (9) minutes in Eastman DK-60 developer at sixty-eight (68) degrees Fahrenheit.

While the operation of the equipment either in the laboratory or in an enclosed gun shed proved entirely satisfactory, it was found impossible to achieve successful synchronization of the unit when used under outdoor conditions on the open ranges. This failure has just recently been traced to the considerable ambient temperature sensitivity of the mercury control thyratron. Excellent synchronization in all cases has since been achieved by the simple expedient of replacing this thyratron by a hydrogen thyratron (type 5C22). A series of tests were recently carried out to determine the accuracy of synchronization of the unit and also its applicability to armor plate investigations. For this purpose, a target consisting of a one-eighth (1/8)-inch sheet of Dural metal was located down range from the gun with its plane at an angle of forty-five (45) degrees to the line of fire. A contact screen consisting of a cardboard sheet on which horizontal conducting bands of aluminum had been sprayed was located approximately eighteen inches behind the target with its plane perpendicular to the line of fire. Connections are made to alternate bands on such a screen so that triggering of the light and camera is effected as soon as the nose of the projectile connects any two bands. Figure 9 shows a picture of a thirty-seven millimeter armor piercing shot which was equipped with a standard, light alloy ballistic cap taken shortly after passage of the projectile through the target plate. It will be noted that the detail (particularly that of the destruction of the ballistic cap) is uniformly good. It will also be observed that triggering has been effected in this instance by the penetration of the inner, armor-piercing cote of the shot through the contact screen. It should be noted here that an Ilex

FIG. 8. Kerr cell photograph of gas cloud surrounding 20-mm shell shortly after complete emergence from the muzzle.

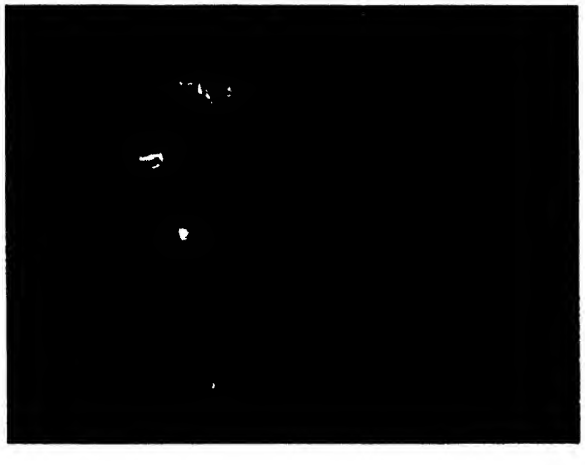


FIG. 9. Kerr cell camera photograph of a thirty-seven millimeter shot taken shortly after penetration of a metal target plate.

"Paragon" lens of aperture $F/4.5$ and focal length six point four (6.4) inches equipped with an electromagnetically operated Acme "Synchro" shutter was employed in this last series of tests; the shutter setting was one twenty-fifth (1/25) second and it was operated via an amplifier by an inertia switch mounted on the barrel of the gun.

ACKNOWLEDGMENTS

The authors wish to record their indebtedness to Mr. E. W. Greenwood of this laboratory for his continued assistance in connection with the purely photographic aspects of the problem and for many invaluable suggestions. Our thanks are also due the Defence Research Board of Canada for permission to publish this report.

Effect of a Circular Groundplane on Antenna Radiation*

A. LEITNER
Mathematics Research Group, Washington Square College, New York University, New York, New York

AND

R. D. SPENCE
Michigan State College, East Lansing, Michigan

(Received May 15, 1950)

The field of a quarter-wave-length antenna above a circular conducting disk of zero thickness—the groundplane—is calculated theoretically by use of the wave functions of the oblate spheroid. Assuming a sinusoidal distribution of current on the antenna, we compute the currents on the groundplane, the radiation resistance and the radiation pattern of the system for various values of the radius a. With these results the distortion of antenna radiation by finite groundplanes can be studied. We find good agreement with recent experimental data.

INTRODUCTION

IT is well known that the behavior of an antenna placed above a conducting plane of either finite or infinite extent differs from that of an identical antenna in free space. While the effect of an infinite plane can easily be predicted in terms of images, the effect of a plane of finite extent poses a much more difficult problem. The latter problem is of some practical importance since in certain situations it may be desirable to place an actual antenna over a conducting surface of finite extent and since in laboratory tests on scale models of antennas it is often assumed that the effect of the earth may be simulated by a finite conducting plate.

Recently Meier and Summers[1] have shown experimentally that the input impedance of a vertical antenna placed above a finite groundplane depends rather markedly on the size of the plane. The only theoretical attack on this problem which has come to the attention of the present authors is the work of Bardeen[2] on the effect of a circular groundplane on a vertical antenna. In the present paper we reconsider essentially the same problem using a method quite different from that of Bardeen. Our method has the advantage that it is less restricted to planes of small radius although even in our case the computation becomes tedious for planes whose radius is very large compared to the length of the antenna.

In the actual problem considered here the groundplane is assumed to be a perfectly conducting disk of radius a and zero thickness. The antenna is assumed to be one quarter of a wave-length long, infinitely thin, and lies along the axis perpendicular to the disk with its lower end in contact with the disk. The antenna is driven in such a way that the current distribution is sinusoidal. While it is fully realized that the last statement postulates a rather idealized situation it hardly seems worth while to take a more elaborate model for a calculation in which the choice of the antenna is not of primary concern.

THE GREEN'S FUNCTION

The central problem of the present calculation consists of constructing a Green's function from which it is possible to calculate the *total* magnetic field **H** at any point in space in terms of the current distribution on the antenna. This means that the Green's function must be so constructed that all the boundary conditions on the electric and magnetic fields **E** and **H** are automatically satisfied on the groundplane. The Green's function will be proper not only in the present problem, but in any circularly symmetric problem involving the perfectly conducting circular disk. Because of the rather peculiar symmetry of the problem, the magnetic field consists of the single azimuthal component H_φ and therefore the Green's function for the magnetic field has a simple scalar form and not the more complicated dyadic form generally found in electromagnetic problems. A further consequence of this symmetry is the fact that H_φ is

* The research reported in this document has been made possible through partial support and sponsorship extended by the Geophysical Research Directorate of the Air Force Cambridge Research Laboratories under Contract No. AF-19(122)-42. It is published for technical information only and does not represent recommendations or conclusions of the sponsoring agency.

[1] A. S. Meier and W. P. Summers, Proc. I. R. E. 37, 609 (1949).
[2] J. Bardeen, Phys. Rev. 36, 1482 (1930).

independent of the azimuthal angle φ. These conditions automatically insure the solenoidal character of **H**. The vector wave equation for **H** reduces to the simple scalar form[3]

$$L[H_\varphi(\rho, z)]=\left(\frac{\partial}{\partial\rho}\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho+\frac{\partial^2}{\partial z^2}+k^2\right)H_\varphi(\rho, z)=0, \quad (1)$$

where ρ and z are the usual cylinder coordinates. These coordinates are not particularly well suited for handling the boundary conditions on the disk and so we introduce a set of oblate spheroidal coordinates η, ξ, φ, which are given by the transformations

$$\rho=a((1-\eta^2)(1+\xi^2))^{\frac{1}{2}}, \quad (2a)$$

$$\varphi=\varphi, \quad (2b)$$

$$z=a\eta\xi, \quad (2c)$$

and for which the surface $\xi=0$ is a disk of zero thickness and radius a. The boundary condition upon the magnetic field on the disk may now be written

$$\partial H_\varphi/\partial\xi=0 \quad \text{when} \quad \xi=0. \quad (3)$$

In the new coordinates Eq. (1) becomes

$$L[H_\varphi(\eta, \xi)]=\left[\frac{\partial}{\partial\eta}(1-\eta^2)\frac{\partial}{\partial\eta}+\frac{\partial}{\partial\xi}(1+\xi^2)\frac{\partial}{\partial\xi}-\left(\frac{1}{1-\eta^2}-\frac{1}{1+\xi^2}\right)+k^2a^2(\eta^2+\xi^2)\right]H_\varphi(\eta, \xi)=0. \quad (4)$$

The solutions of (4) constitute a set of orthogonal functions $\psi_{l1}(\eta, \xi)$ which may be written

$$\psi_{l1}(\eta, \xi)=u_{l1}(\eta)v_{l1}(\xi), \quad l=1, 2, 3\cdots. \quad (5)$$

The functions $u_{l1}(\eta)$ and $v_{l1}(\xi)$ are respectively the "angular" and "radial" oblate spheroidal wave functions[4] of the first order.

The Green's function $G(\eta, \xi; \eta', \xi')$ satisfies

$$L[G(\eta, \xi; \eta', \xi')]=-\frac{\delta(\eta-\eta')\delta(\xi-\xi')\delta(\varphi-\varphi')}{h_\eta h_\xi h_\varphi}, \quad (6)$$

where

$$h_\eta=a((\eta^2+\xi^2)/(1-\eta^2))^{\frac{1}{2}}, \quad h_\xi=a((\eta^2+\xi^2)/(1+\xi^2))^{\frac{1}{2}}, \quad h_\varphi=a((1-\eta^2)(1+\xi^2))^{\frac{1}{2}}, \quad (7)$$

are the metric coefficients of the oblate spheroidal system. Once the Green's function is determined, $H_\varphi(\eta, \xi)$ may be found from

$$H_\varphi(\eta, \xi)=-\int_{\sigma'}[H_\varphi(\eta', \xi')\partial G/\partial n'-G\partial H_\varphi(\eta', \xi')/\partial n']d\sigma' \quad (8)$$

in terms of the values of the integrand on the boundary σ'. The sense of $\partial/\partial n'$ is out of the volume enclosed by σ'. The integral is over the surface of the antenna and the groundplane, provided both H_φ and G satisfy the radiation condition at infinity. If we also require that G satisfy the same boundary condition as H_φ,

$$\partial G/\partial\xi=0 \quad \text{when} \quad \xi=0, \quad (9)$$

then the integrand vanishes on the groundplane.

The conditions imposed upon G suffice to determine it uniquely as a solution of (6), and it will be represented in the wave functions (5).

$$G=\frac{1}{2\pi}\sum_{l=1}^{\infty}\frac{u_{l1}(\eta)u_{l1}(\eta')}{N_{l1}}\begin{cases}{}_1v_{l1}(\xi)^{(3)}v_{l1}(\xi'), & \xi\leq\xi',\\ {}_1v_{l1}(\xi')^{(3)}v_{l1}(\xi), & \xi\geq\xi'.\end{cases} \quad (10)$$

This function is symmetric in η, ξ and η', ξ'. The radial functions $v_{l1}(\xi)$ are solutions of the ordinary differential equation

$$\left[\frac{d}{d\xi}(1+\xi^2)\frac{d}{d\xi}+\frac{1}{1+\xi^2}-\alpha_{l1}+k^2a^2(1+\xi^2)\right]v_{l1}(\xi)=0, \quad (11)$$

which one obtains upon separation of variables in (4). The separation constant is constrained to eigenvalues α_{l1}, $l=1, 2, 3, \cdots$, because of conditions of analyticity in the η dimension. Equation (11) has a fundamental set of solutions, ${}^{(1)}v_{l1}$, ${}^{(2)}v_{l1}$. The linear combination

$${}^{(1)}v_{l1}+i^{(2)}v_{l1}=\underset{\xi\to\infty}{{}^{(3)}v_{l1}\longrightarrow}-(2q_{l1}/ka)e^{ikr}/kr, \quad (12)$$

called *function of the third kind*, has the property that it behaves like an out-going wave at large ξ, as shown in (12).[5] Thus G obeys the radiation condition as ξ or $\xi'\to\infty$, cf. (10). The factor $(1/2\pi)$ normalizes G in the φ dimension and the norm N_{l1} of u_{l1} does the same for η; the quantity q_{l1} is a factor arising from the normalization of the radial functions, discussed in reference 4. Finally, ${}_1v_{l1}$ is the combination

$${}_1v_{l1}=a_l{}^{(1)}v_{l1}+b_l{}^{(2)}v_{l1} \quad (13)$$

chosen such that the condition (9) be satisfied as ξ or $\xi'\to0$, i.e.

$$a_l{}^{(1)}v_{l1}'(0)+b_l{}^{(2)}v_{l1}'(0)=0. \quad (14)$$

A second equation linear in the a_l, b_l is found when (10) is substituted into (6), the result multiplied by any one of the orthogonal set (u_{l1}) and integrated over the range of η, $-1\leq\eta\leq1$. We obtain the differential equation

$$\left[\frac{d}{d\xi}(1+\xi^2)\frac{d}{d\xi}+\frac{1}{1+\xi^2}-\alpha_{l1}+k^2a^2(1+\xi^2)\right]\times\begin{Bmatrix}{}_1v_{l1}(\xi)^{(3)}v_{l1}(\xi')\\ {}_1v_{l1}(\xi')^{(3)}v_{l1}(\xi)\end{Bmatrix}=-\delta(\xi-\xi')/a, \quad \begin{Bmatrix}\xi\leq\xi'\\ \xi\geq\xi'\end{Bmatrix}. \quad (15)$$

[3] Note that neither Eqs. (1) nor (4) represent scalar wave equations.

[4] A comprehensive discussion of these functions in the notation of the present paper may be found in a paper by the present authors. See J. Frank. Inst. 249, 299 (1950).

[5] We assume a dependence $e^{-i\omega t}$ on time.

Integrate this equation with respect to ξ over an interval surrounding the point ξ' and let the interval shrink into this point. The function between braces is continuous but its derivative is by this procedure seen to possess a discontinuity at ξ' of the amount

$$-1/a(1+\xi'^2)={}_1v_{l1}(\xi')^{(3)}v_{l1}'(\xi')-{}_1v_{l1}'(\xi')^{(3)}v_{l1}(\xi'),$$
$$=(ia_l-b_l)W_{l1}({}^{(1)}v_{l1}(\xi'),\ {}^{(2)}v_{l1}(\xi')),$$
$$=(ia_l-b_l)\frac{(-)^{l-1}(2q_{l1}/ka)^2/ka}{1+\xi'^2}, \qquad (16)$$

wherein $W_{l1}({}^{(1)}v_{l1},\ {}^{(2)}v_{l1})$ is the Wronskian of the fundamental set of radial functions whose analytical value is taken from reference 4 and finally substituted above. This leads to a second equation for the constants a_l, b_l and to the result

$${}_1v_{l1}=-\frac{(-)^{l-1}k}{(2q_{l1}/ka)^2}[{}^{(2)}v_{l1}'(0)^{(1)}v_{l1}-{}^{(1)}v_{l1}'(0)^{(2)}v_{l1}]. \qquad (17)$$

Let us call the function inside brackets v_{l1}. Then

$$G=-\frac{k}{2\pi}\sum_{l=1}^{\infty}\frac{(-)^l u_{l1}(\eta)u_{l1}(\eta')}{N_{l1}(2q_{l1}/ka)^2{}^{(3)}v_{l1}'(0)}$$
$$\times\begin{cases}{}_2v_{l1}(\xi)^{(3)}v_{l1}(\xi'), & \xi\leq\xi',\\ {}_2v_{l1}(\xi')^{(3)}v_{l1}(\xi), & \xi\geq\xi'.\end{cases} \qquad (18)$$

On the groundplane $\xi=0$. It can be shown that

$$(-)^{l-1}{}_2v_{l1}(0)/(2q_{l1}/ka)^2=1/ka. \qquad (19)$$

Therefore

$$G(\eta, 0; \eta', \xi')$$
$$=-\frac{1}{2\pi a}\sum_{l=1}^{\infty}u_{l1}(\eta)u_{l1}(\eta')^{(3)}v_{l1}(\xi')/N_{l1}{}^{(3)}v_{l1}'(0). \qquad (20)$$

At large distances $\xi\to\infty$, and

$$G(\eta, \infty; \eta', \xi')$$
$$=\frac{ka}{4\pi}\frac{e^{ikr}}{r}\sum_{l=1}^{\infty}\frac{(-)^{l-1}u_{l1}(\eta)u_{l1}(\eta')_2v_{l1}(\xi')}{N_{l1}q_{l1}{}^{(3)}v_{l1}'(0)}. \qquad (21)$$

DERIVATION OF THE MAGNETIC FIELD

The integral (8) for H_φ is taken over the antenna, a section of the z-axis where $\eta'=1$ and $0\leq\xi'\leq\pi/2ka$, $\partial/\partial n'=\partial/h_{\eta'}\partial\eta'$ and $d\sigma'=(h_{\xi'}h_{\varphi'})_{\eta'=1}d\xi'd\varphi'$. We get

$$H_\varphi=-2\pi a\int_0^{\pi/2ka}[(H_{\varphi'}\partial G/\partial\eta'$$
$$-G\partial H_{\varphi'}/\partial\eta')(1-\eta'^2)]_{\eta'=1}d\xi'. \qquad (22)$$

On the surface of the antenna $H_{\varphi'}$ approaches

$$I_0\cos kz'/2\pi\rho'=I_0\cos ka\xi'/2\pi a(1-\eta'^2)^{\frac{1}{2}}(1+\xi'^2)^{\frac{1}{2}},$$

where $I_0\cos kz'$ is the current on the antenna, and $\partial H_{\varphi'}/\partial\eta'$ approaches

$$I_0\cos ka\xi'/2\pi a(1-\eta'^2)^{\frac{1}{2}}(1+\xi'^2)^{\frac{1}{2}}.$$

When these values and Eqs. (20) and (21) are substituted into the integral one finds

$$H_\varphi(\eta, 0)=-\frac{I_0}{\pi a}\sum_{l=1}^{\infty}{}^{(3)}g_l u_{l1}(\eta)/N_{l1}{}^{(3)}v_{l1}'(0), \qquad (23)$$

$$H_\varphi(\eta, \infty)=\frac{kaI_0}{2\pi}\frac{e^{ikr}}{r}$$
$$\times\sum_{l=1}^{\infty}(-)^{l-1}{}_2g_l u_{l1}(\eta)/N_{l1}q_{l1}{}^{(3)}v_{l1}'(0). \qquad (24)$$

In both expressions

$${}^{(3)}g_l=\int_0^{\pi/2ka}[v_{l1}(\xi')\cos(ka\xi')/(1+\xi'^2)^{\frac{1}{2}}]d\xi'. \qquad (25)$$

The different sub- and superscripts on g_l in (23) and (24) refer to the type of radial function implied in the integrals. These can be evaluated explicitly. We define

$$w_{l1}=(1+\xi^2)^{\frac{1}{2}}v_{l1}, \qquad (26)$$

and find on substitution into (11) that these functions satisfy

$$w_{l1}''+k^2a^2w_{l1}=\alpha_{l1}w_{l1}/(1+\xi^2). \qquad (27)$$

Let

$$y=\cos ka\xi. \qquad (28)$$

Then

$$\left|yw_{l1}'-y'w_{l1}\right|_{\xi_1}^{\xi_2}=\alpha_{l1}\int_{\xi_1}^{\xi_2}[yw_{l1}/(1+\xi^2)]d\xi, \qquad (29)$$

and

$${}^{(3)}g=-{}^{(3)}v_{l1}'(0)[1-ka{}^{(3)}w_{l1}(\pi/2ka)/$$
$${}^{(3)}w_{l1}'(0)]/\alpha_{l1}, \qquad (30)$$

$${}_2g=ka_2w_{l1}(\pi/2ka)/\alpha_{l1}. \qquad (31)$$

Finally

$$H_\varphi(\eta, 0)=\frac{I_0}{\pi a}\sum_{l=1}^{\infty}A_l u_{l1}(\eta), \qquad (32)$$

$$H_\varphi(\eta, \infty)=\frac{I_0k^2a^2}{2\pi}\frac{e^{ikr}}{r}\sum_{l=1}^{\infty}B_l u_{l1}(\eta), \qquad (33)$$

where

$$A_l=\frac{1-ka{}^{(3)}w_{l1}(\pi/2ka)/{}^{(3)}w_{l1}'(0)}{\alpha_{l1}N_{l1}}, \qquad (34)$$

$$B_l=(-1)^{l-1}{}_2w_{l1}(\pi/2ka)/\alpha_{l1}N_{l1}q_{l1}{}^{(3)}v_{l1}'(0). \qquad (35)$$

DISTRIBUTIONS OF CURRENT ON THE GROUNDPLANE

The eigenfunction expansion (32) for the magnetic field on the groundplane is analytic term by term but

the coefficients A_l diverge at least as slowly as the sequence of integers l. We will omit the proof of this assertion here.

The divergence is due to singularities in H_φ which occur when $|\eta| = 1$, that is, when $\rho = a(1-\eta^2)^{\frac{1}{2}} = 0$. These are the points at the centers of both sides of the disk. Now the part of H_φ which is due to the antenna alone has the singularity

$$I_0/4\pi\rho = I_0/4\pi a(1-\eta^2)^{\frac{1}{2}} \text{ as } \rho \to 0.^6 \quad (36)$$

The remaining part of H_φ represents the field scattered by the disk. We will show that it must have a singularity of the form

$$I_0\eta/4\pi a(1-\eta^2)^{\frac{1}{2}} \text{ as } \rho \to 0 \quad (37)$$

on the surface of the disk. The behavior of H_φ would now be

$$H_\varphi(\eta, 0) \to I_0(1+\eta)/4\pi a(1-\eta^2)^{\frac{1}{2}}, \text{ as } \rho \to 0. \quad (38)$$

This will insure continuity between the radial current flowing into the center on the disk and the current flowing upward along the antenna. The latter is $I_0 \cos kz$ and has the value I_0 at the base. The current on the disk (flowing into the center $\eta = +1$ on the top-, and out of the center $\eta = -1$ on the underside) is simply $2\pi\rho H_\varphi(\eta, 0)$. It is clear from Eq. (38) that at $\eta = +1$ the groundplane current will be I_0 and will vanish at $\eta = -1$.

Note also that the scattered part of H_φ is now odd in η, i.e., odd about the plane of the disk. This is a general property in problems involving plane scatterers subject to the boundary condition (3) and can be proven independently.[7]

The even part of the expansion (32) for $H_\varphi(\eta, 0)$ are the terms for which the $(l-1)$ are even. This part corresponds to the value of the magnetic field of the antenna alone on the surface of the disk. There exist explicit expressions for this quantity as already mentioned.[6] The divergence in the even terms is therefore removed when they are replaced by these expressions.

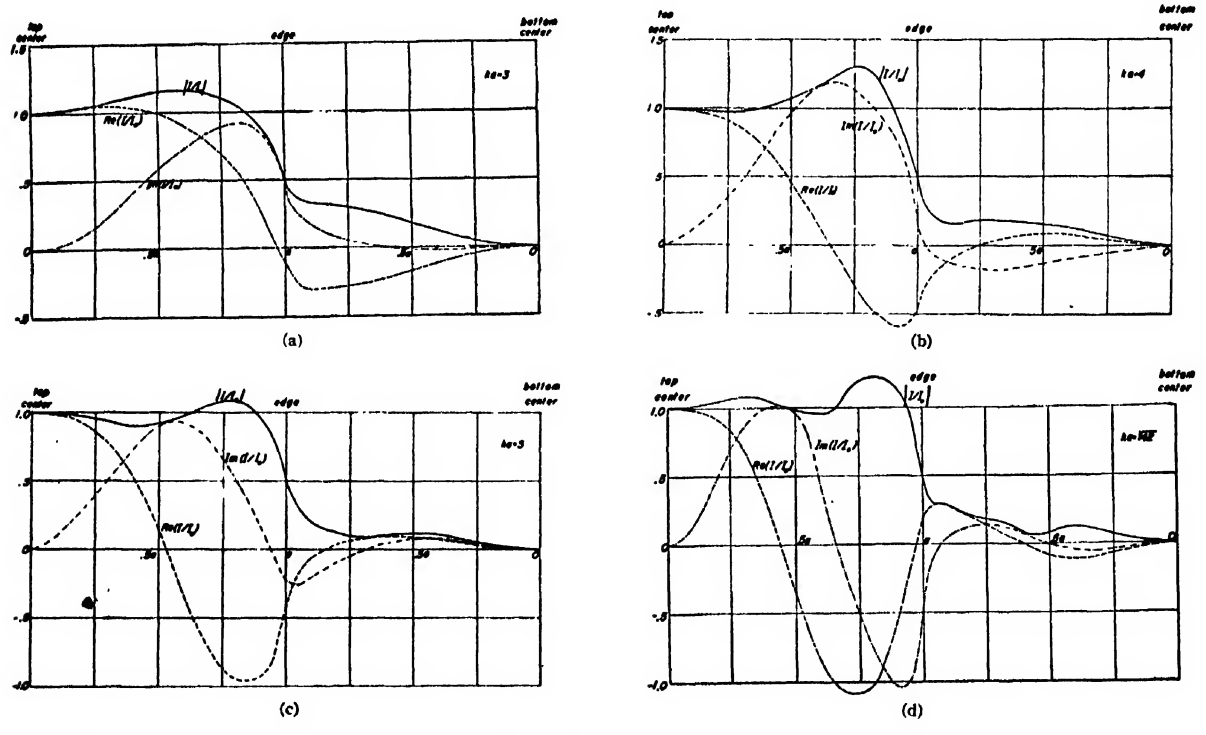
The divergence in the odd part is removed by expanding the singularity in (37) in the odd angular functions,

$$\eta/(1-\eta^2)^{\frac{1}{2}} = \sum_{l=2}{}' C_l u_{l1}(\eta). \quad (39)$$

The series is over alternate integers l (indicated by a prime on the summation sign), and the sequence of coefficients

$$C_l = (1/N_{l1}) \int_{-1}^{1} [\eta u_{l1}(\eta)/(1-\eta^2)^{\frac{1}{2}}] d\eta, \ (l-1) \text{ odd}, \quad (40)$$

must necessarily diverge.



FIGS. 1(a) to (d). Ratio of radial current I (on both sides of the groundplane) to the amplitude I_0 of the current on the antenna. Plotted against distance from the center for the cases when $ka = 3, 4, 5, (42)^{\frac{1}{2}}$.

[6] One may obtain this result upon substitution into the expressions for the field of our antenna which are given in the literature. See for instance J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), p. 457, Eq. (75). (Practical units are used in the present paper.)

[7] See, for example, A. Leitner, J. Acous. Soc. Am. 21, 331 (1949).

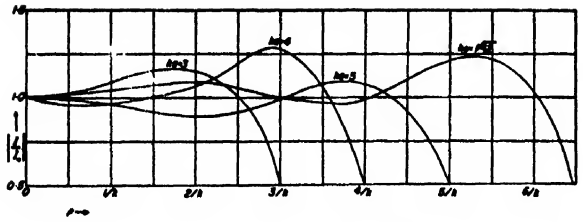


FIG. 2. Absolute values of the ratio of radial current I (on the *antenna* side of the groundplane) to the amplitude I_0 of the current on the antenna. Plotted to common scale against distance from the center for the cases when $ka=3, 4, 5, (42)^{\frac{1}{2}}$.

Finally, the current $I(\eta)$ on the groundplane has the value

$$I(\eta)=2\pi\rho H_\varphi(\eta,0)$$
$$=(I_0/2)[-i\exp[ik(\rho^2+(\lambda/4)^2)^{\frac{1}{2}}]$$
$$+\eta+(1-\eta^2)^{\frac{1}{2}}\cdot\sum_{l=2}^{\infty}{}'(4A_l-C_l)u_{l1}(\eta)]. \quad (41)$$

Numerical computations show that the remaining series in this expression converges.

The ratios of $I(\eta)$ to I_0 are plotted against distance from the base of the antenna in Fig. 1. Since $I(\eta)$ is continuous at the edge $\rho=a$, it was possible to plot the values on both sides continuously on one graph.

As ka increases, the currents on the underside become smaller, while on the antenna side the average of the absolute value becomes more nearly unity for each ka. Figure 2 illustrates this latter point more clearly. Theer we have plotted the currents on the *antenna* side on one graph for the various values of the parameter ka, with a common scale for the distance from the center. The larger ka correspond to larger groundplanes.

The trends revealed in the figures are consistent with the case of an infinite groundplane. In that case no current would flow on the underside, while the absolute value would be I_0 at all points on the upper side.

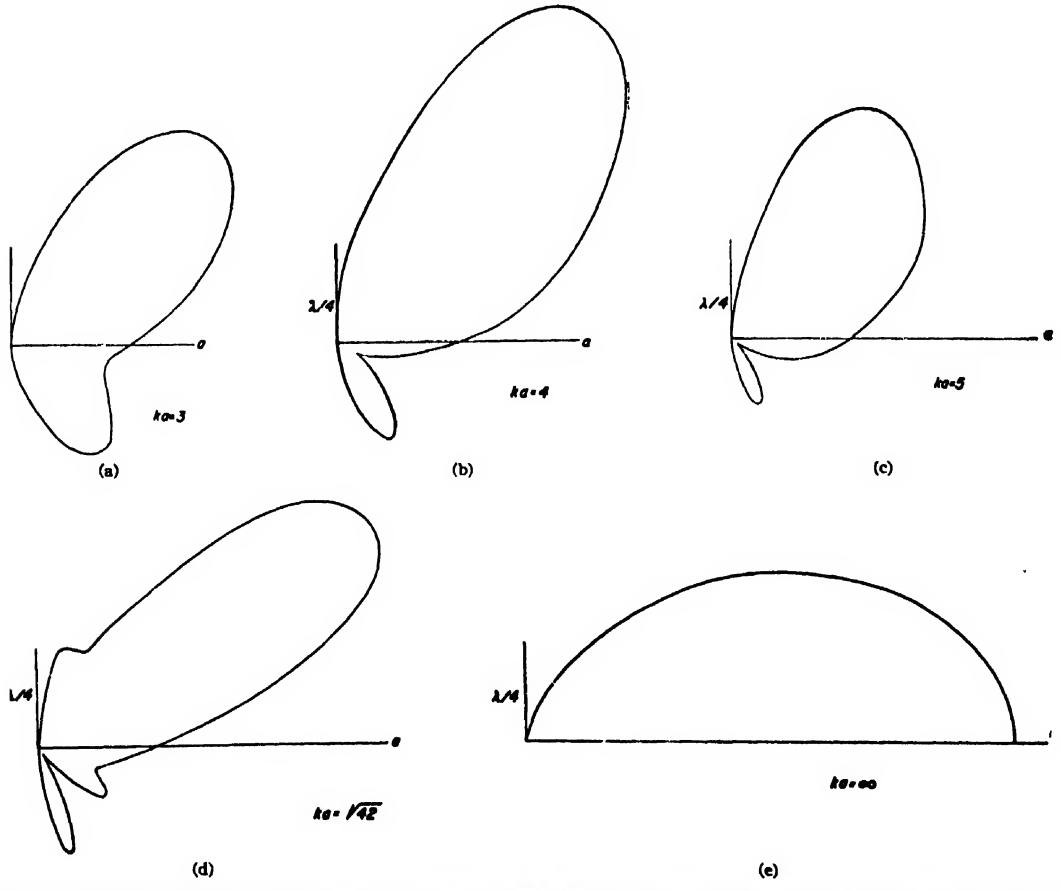
RADIATION PATTERNS

We next study the field at large distances from the radiating system.

As $r\to\infty$ the electric vector is entirely transverse to the radial direction,

$$E\to E_\eta\to(\mu_0/\epsilon_0)^{\frac{1}{2}}H_\varphi, \quad (42)$$

where μ_0 and ϵ_0 are the free space permeability and dielectric constant, respectively. So the Poynting flux



FIGS. 3(a) to (e). Radiation patterns (polar graphs in any azimuthal plane) of a quarter wave-length antenna situated above circular groundplanes for the cases when $ka=3, 4, 5, (42)^{\frac{1}{2}}, \infty$.

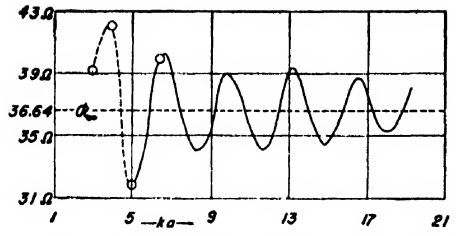


FIG. 4. Radiation resistance of the system of antenna and groundplane plotted against ka. *Solid curve:* experimental data of Meier and Summers. *Dotted curve:* extrapolation of data. *Circles:* calculated values. *Dotted horizontal line:* corresponds to $ka=\infty$.

becomes

$$S=\tfrac{1}{2}E_\eta\tilde{H}_\varphi$$

$$=(I_0^2/8\pi r^2)(\mu_0/\epsilon_0)^{\frac{1}{2}}(ka)^4\Big|\sum_{l=1}^{\infty}B_l u_{l1}(\eta)\Big|^2. \quad (43)$$

This quantity is to be compared to the case of the infinite groundplane, when the Poynting flux is the same as for the half-wave-length antenna in the region $\theta\leq\pi/2$, and vanishes below. In this familiar case[8]

$$S_\infty=(I_0^2/8\pi r^2)(\mu_0/\epsilon_0)^{\frac{1}{2}}\cos^2[\pi(\cos\theta)/2]/\sin^2\theta.^9 \quad (44)$$

(The subscript ∞ refers to the infinite groundplane.)

A polar graph of the dependence of S on the angle θ is the radiation pattern. It is drawn, to common scale, for each of the parameter values $ka=3, 4, 5, (42)^{\frac{1}{2}}$ and ∞, in Fig. 3.

We see that the patterns 3(a) to (d) differ greatly from the pattern 3(e). The radiation field shows no sign of approaching the case corresponding to the isolated half-wave-length antenna. Technical applications may possibly be found for the peculiar directivity of the finite groundplane system illustrated by the figure: while the isolated antenna radiates its energy most strongly in a direction perpendicular to the antenna, the finite groundplane system radiates preponderantly in an oblique direction.

RADIATION RESISTANCE

The radiation resistance of the finite groundplane system is calculated from the Poynting flux, and becomes

$$\mathfrak{R}=(1/2\pi)(\mu_0/\epsilon_0)^{\frac{1}{2}}(ka)^4\sum_{l=1}^{\infty}N_{l1}|B_l|^2\ \text{ohms}, \quad (45)$$

while in the infinite case

$$\mathfrak{R}_\infty=36.64\ \text{ohms}, \quad (46)$$

exactly one-half the resistance of the half-wave antenna.

These values are plotted in Fig. 4. The solid line represents the experimental data of Meier and Summers, taken on thin wire antennas approximately one quarter of a wave-length long with thin circular conducting plates as groundplanes. The results of our calculation are entered as circles and lie on the experimental curve or an extrapolation of this curve, drawn as a dotted line.

The value of $\mathfrak{R}$ never differs greatly from $\mathfrak{R}_\infty$ and, as a function of ka, oscillates about $\mathfrak{R}_\infty$, approaching this value asymptotically, as is to be expected when $ka\to\infty$.

When one wishes to measure the resistance of an isolated antenna, the rule has been to mount the test model over a groundplane, and consider the resistance of the system as the resistance of the isolated antenna. The error in this measurement depends on the relative size of the groundplane, as follows clearly from the curve in Fig. 4. An improvement over this procedure suggests itself from the nature of the curve. Rather than a single groundplane, one might use a whole series of them, similar in shape but differing in size, for each single antenna to be tested. A graph of resistance against the size of the groundplane will resemble the curve in Fig. 4. From such a curve it should be possible to estimate the position of the asymptote $\mathfrak{R}=\mathfrak{R}_\infty$ with greater accuracy.

We wish to thank Professor C. Kirkuchi for reading the manuscript and are particularly grateful to Miss J. Bruno for performing the computations.

[8] See reference 6, p. 441, Eq. (11).
[9] Note that $\eta\to\cos\theta$ as $r\to\infty$.

Turbulence in Apparatus for Measurement of Streaming Double Refraction

H. G. JERRARD
Physics Department, University College, Southampton, England
(Received March 16, 1950)

The different apparatus used by some thirty investigators, who have induced double refraction in a liquid contained in the annular gap between two concentric cylinders, are examined with particular reference to the onset of turbulence. In the case where the outer cylinder alone rotates, curves are given which easily enable the critical gradients corresponding to the dimensions of the apparatus used to be found. The results of the examination are so tabulated that present and future workers in this field can quickly decide on the significance of the measurements obtained by any of the investigators. An explanation is put forward of the surprising fact that with either the inner or outer cylinder rotating measurements of double refraction apparently show no very sharp break between regions of laminar and turbulent flow. The value of making simultaneous measurements of streaming double refraction and viscosity is pointed out and the advantages of using an apparatus with rotating outer and stationary inner cylinder are briefly summarized. Mention is made of optical design.

I. INTRODUCTION

IT was observed by Maxwell[1] that Canada balsam, although normally isotropic, becomes birefringent when subjected to a shearing stress. The name streaming double refraction is usually applied to this effect in analogy to the terms magnetic and electrical double refraction. Since the original work of Maxwell, streaming double refraction in liquids has been extensively studied and is now recognized as a powerful method for the determination of molecular size and shape.[2]

The essential condition for the appearance of streaming double refraction is the production of a velocity gradient in the moving liquid, which is then examined by a suitable optical method. As a prelude to the construction of apparatus for streaming double refraction experiments, the descriptions of the apparatus used by previous workers in this field have been examined. It appears that only two methods are in use to produce the required velocity gradient. One method is to produce a gradient by flow of a liquid in a narrow tube of rectangular cross section and the other by motion of a liquid confined in the annular gap between two concentric cylinders, one of which is stationary, the other rotating.

The first method is suitable for the production of gradients of the order of 2000–3000 sec.$^{-1}$ but which, however, are not constant across the tube. For pure liquids the gradient varies linearly from zero at the center to a high value near the walls; this corresponds to a parabolic distribution of velocity across the tube. Colloidal solutions on the other hand show a non-linear variation of gradient from zero at the center to a maximum at the walls and give a non-parabolic distribution of velocity.[3]

In both cases, therefore, there is no uniquely defined value of the velocity gradient which can be correlated with a given double refraction measurement. The second method is capable not only of producing an almost constant velocity gradient across the gap for both homogeneous and colloidal solutions but enables very high gradients (of the order of 30,000 sec.$^{-1}$) to be obtained.

The majority of observers have used this second method but unfortunately many workers have paid insufficient attention to the possibility of turbulence. The latter is, however, of prime importance since streaming double refraction is not amenable to theoretical treatment in the turbulent region and the results obtained for the double refraction, although of the right order of magnitude, can have no quantitative significance. It is a surprising fact, however, that except in a few isolated cases of colloidal solutions the onset of turbulence does not seem to produce any sharp break in curves showing the variation of double refraction with velocity gradient.[4]

In this paper the results of the examination, from the point of view of turbulence, of the concentric cylinder apparatus described by various authors are presented and reasons why the onset of turbulence may escape notice are given. It is hoped that the results may be of use to other workers in the fields of streaming double refraction and viscometry.

II. THEORY

In the concentric cylinder method, either the inner cylinder can rotate with the outer stationary or vice versa. The equations of motion for steady flow are the same in both cases but there is considerable difference in the velocity at which turbulence first occurs, because the distribution of centrifugal forces for the case of the rotating outer cylinder is such as to stabilize the motion of the fluid, whereas with the rotating inner cylinder this is not so.

1. Steady Motion

Suppose V be the velocity at any point P of an incompressible viscous fluid in steady motion between

[1] J. C. Maxwell, *Scientific Papers* (Cambridge University Press, London, 1890), Vol. II, p. 379. Proc. Roy. Soc. A22, 46 (1873).
[2] See review by J. T. Edsall, *Advances in Colloid Science* (Interscience Publishers, Inc., New York, 1942), Vol. 1, 269.
[3] A. S. C. Lawrence, Proc. Roy. Soc. A148, 59 (1935); W. Philippoff, Kolloid. Zeits. 75, 142 (1936).
[4] C. Sadron, J. Phys. Radium 7, 263 (1936); Schweiz. Arch. angew. Wiss. Tech. (3–4) 158, 8 (1937).

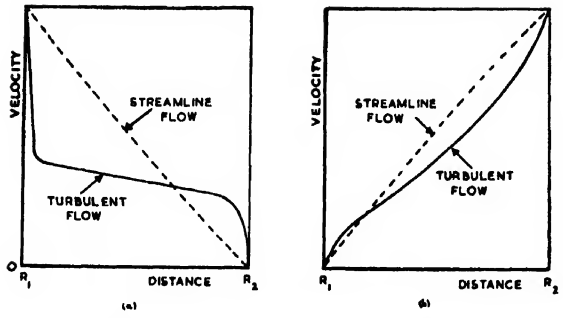


FIG. 1. Velocity distribution across the annular gap. (a) Inner cylinder rotating (after Wattendorf). (b) Outer cylinder rotating (after G. I. Taylor).

two infinitely long concentric rotating cylinders of radii R_1 and R_2, $(R_2>R_1)$. If r is the distance of P from the axis and if Ω_1 and Ω_2 are the angular velocities of the cylinders, then from the Navier-Stokes equations,[5] it may be shown that:

$$V=Ar+(B/r),$$

where

$$A=\frac{R_2^2\Omega_2-R_1^2\Omega_1}{R_2^2-R_1^2} \quad \text{and} \quad B=\frac{R_1^2R_2^2(\Omega_1-\Omega_2)}{R_2^2-R_1^2}.$$

The linear velocity gradient G is given by

$$G=dV/dr=A-(B/r^2)$$

and the change of gradient with distance is

$$dG/dr=2B/r^3.$$

If either cylinder is stationary and the other has an angular velocity Ω, then for small gap widths, the change in gradient across the gap is given by $dG=\Omega$, and the average gradient by $G=\Omega R/s$. In the latter $R=R_2\doteq R_1$, $s=R_2-R_1$ and $s\ll R$. Hence, $dG/G=s/R$. This ratio may be taken as representative of the constancy of the velocity gradient: it will be called the constancy factor.[6] Clearly if s/R is small, then so also is the error in assuming the gradient to have the constant value $(R\Omega)/s$ across the gap. The importance of this will be seen later.

2. Turbulent Motion

The stability for symmetrical disturbances of a viscous liquid in steady motion between concentric cylinders has been the subject of a very thorough theoretical and experimental treatment by Taylor.[7]

(a) *Inner cylinder rotating.* The motion of the liquid in this case is two-dimensional and according to Taylor is stable for angular velocities of the cylinder not exceeding a critical value Ω_c, given by the equation

$$\Omega_c^2=\frac{\pi^4\nu^2(R_1+R_2)}{2Ps^3R_1^2}, \tag{1}$$

which corresponds to a critical Reynolds number Re such that

$$\mathrm{Re}=\frac{(R_1\Omega_c)s}{\nu}=\pi^2\left(\frac{R_1+R_2}{2Ps}\right)^{\frac{1}{2}}. \tag{2}$$

In Eqs. (1) and (2), ν is the kinematic viscosity, s, R_1 and R_2 have the same meaning as before and P is a constant which Taylor gives as

$$P=0.0571[1-0.652(s/R_1)]+0.00056[1-0.652(s/R_1)]^{-1}.$$

When s/R_1 is small, P may be taken as 0.057. If s/R_1 is not small the value of P above, which neglects second-order terms of s/R_1 and Taylor states is to an order of approximation of about 1 percent, holds when $s/R_1\geqslant 0.33$. Some other limitations of Taylor's criteria are mentioned later. The validity of Eq. (1) was investigated experimentally by Lewis,[8] who states that it holds for values of s/R_1 as high as 0.71 and for liquids whose coefficients of kinematic viscosity vary from 0.006 to 0.018 stokes. A less detailed theoretical study, which has the merit that it involves no approximations based on the thinness of the liquid layer, has been given by Synge.[9]

As soon as the velocity given by Eq. (1) is exceeded the laminar motion is succeeded by a three-dimensional motion such that the liquid breaks up into symmetrical ring-shaped vortices spread at regular intervals along the length of the cylinder: these vortices rotate alternately in opposite directions. It is true that this is not exactly turbulence, as we generally understand it, since the pattern is regular and stationary, but it is an intermediate stage between laminar and turbulent motion.

In an experimental investigation of the velocity

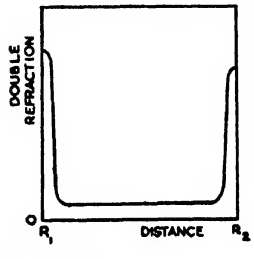


FIG. 2. Distribution of double refraction across annular gap for case of rotating inner cylinder.

[5] See for example H. Lamb, *Hydrodynamics* (Cambridge University Press, London, 1930), 5th Ed. Chapter 11.

[6] For the dimensions of the apparatus given in Tables I and II, the radii of the outer and inner cylinders are not always approximately equal. In order that the values of the constancy factor shall be comparable, the inner cylinder radius, R_1, has been used instead of R.

[7] G. I. Taylor, Phil. Trans. A223, 289 (1923); G. I. Taylor, Proc. Roy. Soc. A102, 541 (1923); G. I. Taylor, Proc. Roy. Soc. A157, 546 (1936).

[8] J. W. Lewis, Proc. Roy. Soc. A117, 388 (1928). For a description of the apparatus used by Lewis see J. W. Lewis and E. N. da C. Andrade, J. Sci. Inst. 1, 373 (1923).

[9] J. L. Synge, Proc. Roy. Soc. A167, 250 (1938).

distribution in the gap, Taylor[10] showed that there was a large change in velocity near the walls but throughout some 83 percent of the gap the product of linear velocity and radius was constant. This confirmed the work of Wattendorf[11] who obtained a velocity distribution curve of the type shown in Fig. 1a. These facts can be explained on the theory that the dynamics of turbulent motion should be regarded as a transport of vorticity[12] and is important because it explains the phenomenon mentioned by Snellman[13] and Buchheim, Stuart, and Menz[14] that in some cases, near the surfaces of the cylinders doubly refracting layers have been seen and between them a domain without any double refraction.[15] Thus if V_1 and V_2 be the velocities at points of the liquid

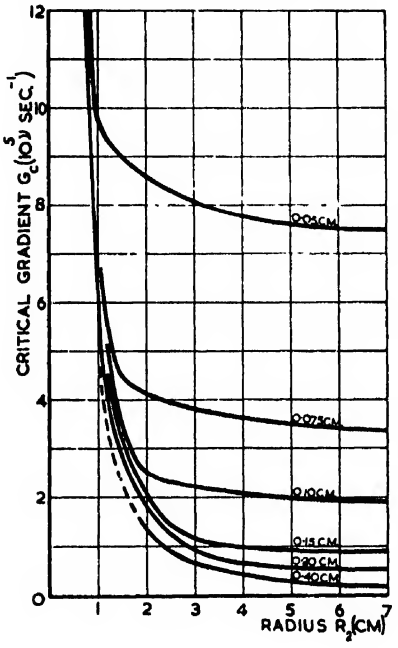


FIG. 3. Variation of critical velocity gradient with radius of outer cylinder. Gap widths 0.05, 0.075, 0.10, 0.15, 0.20, 0.40 cm. Outer cylinder rotating.

distance r and $r+x$, respectively from the axis, then

$$V_1 r = V_2(r+x) = K \text{ (a constant)},$$

so that

$$V_2 - V_1 = Kx/r(r+x) \qquad (3)$$

or the linear velocity gradient, $(V_2-V_1)/x = K/(r(r+x))$. The gradient, therefore, is approximately inversely proportional to r^2 and will be small at the center of the gap where Eq. (3) holds and large near the surfaces of

[10] G. I. Taylor, Proc. Roy. Soc. A151, 494 (1935).
[11] F. L. Wattendorf, Proc. Roy. Soc. A148, 565 (1935).
[12] G. I. Taylor, Proc. Roy. Soc. A135, 685 (1932).
[13] O. Snellman, Ark. Kemi. Min. Geol. 19A, 30, 3 (1945).
[14] Buchheim, Stuart, and Menz, Zeits. f. Physik 112 (7–8), 407 (1939).
[15] In a private communication, Dr. Snellman states his was an experimental observation with a rotating inner and stationary outer cylinder.

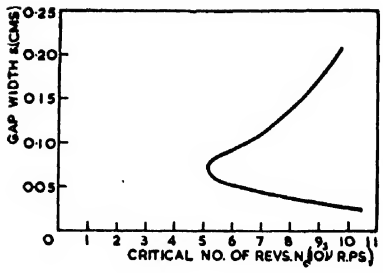


FIG. 4. Variation of critical number of r.p.s. with width of gap. Outer cylinder radius = 1.25 cm.

the cylinders where the velocity change is considerable (see Fig. 1a). Doubly refracting layers will, therefore, be more easily seen at the cylinder surfaces, the distribution of double refraction across the gap being as in Fig. 2. If a well-aligned optical system is used for observation such a distribution may be detected,[16] but with a less efficient system the double refraction may appear as a mean value, uniform across the gap. The mean value may be greater than that which would exist for laminar flow if the layers near the cylinder surfaces are traversed by the light and smaller (even zero) if the middle region is traversed. Thus as previously mentioned the onset of turbulence may cause no sharp break in the curve showing variation of double refraction with gradient but perhaps a slight departure from the curve for laminar flow to give either too high or too low values of the double refraction. This has been detected by Sadron.[17]

(b) *Outer cylinder rotating.* In this case the critical velocity is very much higher than that when the inner cylinder is the rotor. Taylor[18] gives the value of Ω_c in the form of a curve which shows $\log_{10}(\Omega R_2 s/\nu)$ plotted against $\log_{10}(s/R_2)$. This curve is good for Taylor's purpose, but is somewhat tedious to use if one wishes to find quickly the critical speeds and gradients for a specific piece of apparatus. For this reason curves shown in Fig. 3 have been drawn from data derived from Fig. 11 of Taylor's article: they show the values of the critical gradients plotted against the radii of the outer cylinder for various gap widths. Figure 3 must be understood as giving a limit to the critical gradient below which no turbulence occurs. Because of the inevitable inaccuracy in obtaining data from Taylor's published diagram, Fig. 3 does not give with equal certainty the upper limits. These curves should prove of value in the design and use of high speed concentric cylinder viscometers.

From these curves it is easy to calculate the critical number of revolutions per sec. N_c for any system (i.e. the number of revolutions per sec. above which turbulence occurs) and a graph of s against N_c for a fixed value of the radius takes the form shown in Fig. 4. This

[16] In the experiments of Sadron on pure liquids and colloidal solutions (Schweiz. Arch. angew. Wiss. Tech. (3–4) 158, 8 (1937–(38)) this appears as a non-uniform illumination of the half-shadow.
[17] C. Sadron, J. Phys. Radium 7, 263 (1936).
[18] G. I. Taylor, Proc. Roy. Soc. A157, 546 (1936).

TABLE I. Dimensions of cylinders in apparatus used by various investigators and critical velocity gradients—inner cylinder rotating.

Author		Inner cylinder rad. (R_1) cm	Gap width (s) cm	Gradient constancy factor $\frac{s}{R_1}\frac{dG}{G}=-(10^{-3})$	Critical angular velocity Ω_c (ν-rad. sec.$^{-1}$)	Critical velocity gradient G_c (ν sec.$^{-1}$)	Critical No. of revs. N_c (ν r.p.s.)
Kundt[a]	(a)	0.95	0.25	26.3	393	1493	63
	(b)	2.50	0.40	16.0	113	706	18
	(c)	0.95	0.10	10.5	1420	13490	226
de Metz[b]	(a)	1.75	0.50	28.6	104	363	16
	(b)	1.75	1.00	57.1	44	77	7
Umlauf[c]	(a)	2.33	0.37	15.9	131	825	21
	(b)	2.10	0.60	28.6	72	252	11
Krueger[d]		2.00	0.25	12.5	250	2000	40
Vorlander and Walter[e]		1.05	0.10	9.5	1340	14070	213
Vorlander and Kirchner[f]		1.10	0.10	9.1	1330	14630	212
Signer[g, h]		2.300	0.195	8.5	330	3897	52
Boeder[i]	(a)	4.05	0.20	4.9	235	4755	37
	(b)	4.05	0.10	2.5	655	26520	104
Vorlander and Fisher[j, k]		9.20	0.05	0.54	1215	223600	193
Signer and Gross[l]		2.4700	0.0252	1.0	6576	644500	1047
Sadron[m, n*]	(a)	2.498	0.052	2.1	2220	106600	353
	(b)	2.420	0.130	5.4	581	10680	91
	(c)	2.500	0.050	2.0	2350	117500	374
	(d)	2.528	0.022	0.87	7960	914700	1267
Nitschmann[o]		2.475	0.0427	1.7	2990	173300	476
Buchheim,† Stuart and Menz[p]	(a)	4.991	0.039	0.78	2400	307100	382
	(b)	4.871	0.159	3.3	299	9160	47
	(c)	4.770	0.260	5.45	146	2678	23
Nitschmann and Guggisberg[q]		2.4800	0.0209	0.84	8680	1030000	1381
Snellman and Bjornstahl[r, s]	(a)	3.50	0.05	1.43	1981	138700	315
	(b)	3.45	0.10	2.90	711	24550	113
	(c)	3.40	0.15	4.41	393	8910	63
de Rosset[t]		5.266	0.0268	0.51	4094	804400	652
Edsall[u, v] *et al.*		2.512	0.025	0.99$_5$	6598	663000	1050
Tsvetkov[w] *et al.*		Data published insufficient to enable critical values to be calculated.					

[a] A. Kundt, Wied. Ann. 13, 110 (1881).
[b] G. de Metz, Ann. Phys. (Leipzig) 35, 497 (1888).
[c] K. Umlauf, Ann. Phys. (Leipzig) 45, 304 (1892).
[d] E. Kreuger, Zeits. f. physik. Chemie A109, 438 (1924).
[e] D Vorländer and R. Walter, Zeits. f. physik. Chemie A118, 1 (1925).
[f] D. Vorländer and U. Kirchner, Zeits. f. physik. Chemie A152, 47 (1930).
[g] R. Signer, Zeits. f. physik. Chemie A150, 257 (1930).
[h] This apparatus was also used by G. Boehm and R. Signer, Helv. Chim. Acta. 14, 1370 (1931).
[i] P. Boeder, Zeits. f. Physik 75 (3-4), 258 (1932).
[j] D. Vorländer and J. Fischer, Berichte 65B, 1756 (1932).
[k] This apparatus was also used by D. Vorländer and P. Specht, Zeits. f. physik. Chemie A178, 93 (1936).
[l] R. Signer and H. Gross, Zeits. f. physik. Chemie A165, 161 (1933).
[m] C. Sadron, C.R. Acad. Sci. (Paris) 202, 404 (1936); Schweiz. Arch. angew. Wiss. Tech. (3-4) 158, 8 (1937); J. Phys. Radium 7, 263 (1936).
[n] See also C. Sadron and H. Mosimann, J. Phys. Radium. 9, 384 (1938); Sadron, Bonot, and Mosimann, J. Chim. Phys. 36, 78 (1939).
* Although Taylor's formula as quoted by Sadron is incorrect it seems that this is a typographical error as the correct formula appears to have been used to calculate Ω_c and G_c.
[o] H. Nitschmann, Helv. Chim. Acta 21, 315 (1938).
† These authors appear to have made a numerical mistake in copying Taylor's formula. Their published values of N_c must be divided by $\sqrt{2}$.
[p] Buchheim, Stuart, and Menz, Zeits. f. Physik 112 (7-8), 407 (1939).
[q] H. Nitschmann and H. Guggisberg, Helv. Chim. Acta 24, 434, 574 (1941).
[r] O. Snellman and Y. Bjornstahl, Kolloid. Beih. 52, 403 (1941).
[s] This apparatus was also used by O. Snellman, Ark. Kemi. Min. Geol. B19 (No. 5), 1 (1944); Acta. Chem. Scand. 1, 291 (1947).
[t] A. I. de Rossett, J. Chem. Phys. 9, 766 (1941).
[u] Edsall, Gordon, Mehl, Scheinberg, and Mann, Rev. Sci. Inst. 15, 243 (1944).
[v] This apparatus was also used by J. F. Foster and J. T. Edsall, J. Am. Chem. Soc. 67, 617 (1945); Edsall, Foster, and Scheinberg, J. Am. Chem. Soc. 69, 2731 (1947); J. T. Edsall and J. F. Foster, J. Am. Chem. Soc. 70, 1860 (1948).
[w] V. N. Tsvetkov and A. Petrova, J. Tech. Phys. U.R.S.S. 12, 423 (1942); V. N. Tsvetkov and E. Frisman, Acta. Physicochim. U.R.S.S. 20, 61 (1945). The published details of the apparatus give only the gap widths and so it is impossible to calculate Ω_c and G_c. However, the formula used by these workers for calculating these values is identical with that used by Buchheim, Stuart and Menz and is incorrect. The anomalies mentioned may therefore be partly due to turbulence.

curve clearly shows that as s increases from zero to infinity so N_c changes from infinity, through a minimum value to a large value again, so that for a given radius of the outer cylinder a decrease in gap width may or may not allow an increase in the speed of revolution. Beyond a certain value of s, as s increases, so N_c increases but the advantage derived by having a large gap is counteracted by the fact that the gradient constancy factor increases, which means that the error in assuming the gradient to be constant across the gap is now larger.

Taylor[19] investigated the velocity variation across the gap and found (see Fig. 1b) that (a) the velocity distribution is similar to that obtained with stream line flow, (b) turbulence is not very effective at the outer layers and (c) due to the stabilizing effect of rotation there is a large velocity gradient across the gap. Consideration of Fig. 1b indicates that even when turbulence sets in a sharp break may not occur in the double refraction velocity gradient curves as the gradient does not differ greatly from that occurring with stream-line flow.

[19] G. I. Taylor, Proc. Roy. Soc. A157, 565 (1936).

III. APPLICATION TO EXISTING APPARATUS

(a) Inner Cylinder Rotating

The criterion for turbulence given by Eq. (1) has been applied to the apparatus of the majority of observers who have published the dimensions used. The results are summarized in Table I. Only in the apparatus of de Metz does the ratio s/R_1 exceed 0.33 and even then it is less than 0.71. The values of Ω_c, G_c, and N_c given,

can in all cases therefore, be taken as accurate to 1 percent. To find which of the results of the various authors are taken for conditions of laminar flow it is only necessary to calculate the value of ν for the liquid used. If then the number of revolutions per second of the cylinder exceeds that given in the seventh column, the results cannot be regarded as being of any quantitative significance.

(b) Outer Cylinder Rotating

The results in this case are given in Table II in which the values quoted for the critical constants are lower limits; in fact values given for $s/R_2 > 0.22$ are probably very much below the truth. It is immediately obvious that the values of G_c are very much greater than in the case of rotation of the inner cylinder although the constancy factor is greater. In only one case were the measurements made in the turbulent region and that is in the work of Winkler who studied benzene at gradients of 7710, 11,600 and 14,000 sec.$^{-1}$. If ν is assumed to be 0.0074 stokes[20] at 20°C the critical gradient is about 420 sec.$^{-1}$. Winkler was led to believe he was working in the laminar region for he refers only to Taylor's conclusion published in 1923[21] that flow should be laminar at all speeds of rotation. This conclusion was superseded by Taylor's 1936 paper,[22] to which Winkler made no reference.

IV. OPTICAL DESIGN

The concentric cylinders form only one component of the apparatus for investigating streaming double refraction. In designing an apparatus attention must also be paid to the optical components, for there are numerous optical sources of error, which, if ignored, are liable to introduce double refraction of the same order as that being measured. The main errors are (a) polarization at lenses, (b) polarization by reflection from metal surfaces (c) strain in the cover glasses through which the light leaves the liquid (d) the effect of heat developed in the liquid annulus.

An excellent survey of these effects has been given by Snellman[23] and they will, therefore, not be discussed in detail here. It is sufficient to say that no trouble occurs from (a) if no lenses are situated between the polarizing prisms; the effects of (b) are avoided if the optical system as developed by Bjornstahl[24,25] be used; the effects of (c) are avoided if the method of Frey-Wyssling and Weber[26] be used while the effect of the heat developed can be calculated by a theory developed by Bjornstahl.[27]

V. GENERAL REMARKS

From the results obtained in streaming double refraction experiments it is possible to get information on the properties of the particles or molecules in the liquid. One of the important parameters required to interpret the results is the viscosity under the prevailing conditions. The fact that the viscosity coefficient must be that measured under the prevailing conditions is especially important for non-Newtonian liquids or those exhibiting anomalous viscosity. Results derived from a combination

TABLE II. Dimensions of cylinders in apparatus used by various investigators and critical velocity gradients—outer cylinder rotating.

Author		Inner cylinder rad. (R_1) cm	Gap width (s) cm	Gradient constancy factor $\frac{s}{R_1}\frac{dG}{G}=-(10^{-2})$	Critical angular velocity Ω_c (ν-rad. sec.$^{-1}$)	Critical velocity gradient G_c (ν sec.$^{-1}$)	Critical No. of revs. N_c (ν r.p.s.)
Freundlich, Stapelfeldt and Zocher[a]	(a)	1.00	0.45	45.0	69000	222000	11000
	(b)	0.35	1.10	314.3	28000*	37000	4400
	(c)	0.55	0.90	163.6	34500*	55500	5500
von Muralt and Edsall[b]	(a)	0.91	1.09	119.8	21000*	38000	3300
	(b)	1.30	0.70	53.8	32000*	92000	5100
	(c)	1.80	0.20	11.1	17300	173000	2750
Robinson[c]		1 7275	1.135	65.7	13800*	35000	2200
Winkler[d]		5.06	0.2	3.9$_5$	2130	56000	340
Lawrence†[e] *et al.*	(a)	0.20	0.192	96.0	600000*	1221000	95500
	(b)	0.32	0.072	22.5	707000	3849000	112500

* In these cases $s/R_2 > 0.22$: the critical values are probably very much higher than those given here.

[a] Freundlich, Stapelfeldt, and Zocher, Zeits. f. physik. Chemie 114, 161, 190 (1925).

[b] A. L. von Muralt and J. T. Edsall, J. Biol. Chem. 89, 351 (1930); Trans. Faraday Soc. 26, 837 (1930).

[c] J. R. Robinson, Proc. Roy. Soc. A170, 519 (1939).

[d] E. Winkler, Zeits. f. Physik 118 (3-4), 232 (1941). See also E. Winkle and W. Kast, Naturwiss. 29, 288 (1941).

† Some uncertainty exists as to the actual dimensions of this instrument. It would appear from the values of the gap width given by the authors that the figures given for the radii are actually those for the diameter. In any case there can be no question of turbulence occurring at the speeds used.

[e] Lawrence, Needham, and Shen, J. Gen. Physiol. 27, 201 (1944).

[20] This value is one derived from measurements with a flow tube viscometer.

[21] G. I. Taylor, Phil. Trans. A223, 289 (1923).

[22] G. I. Taylor, Proc. Roy. Soc. A157, 546 (1936).

[23] O. Snellman, Ark. Kemi. Min. Geol. 19A (30), 1 (1945).

[24] Y. Bjornstahl, J. Opt. Soc. Am. 29, 201 (1939).

[25] Frey-Wyssling and Weber (Helv. Chim. Acta 24, 278 (1949) suggested a vectorial method of correcting this effect but their method besides being laborious may not be valid owing to the assumptions used. If Bjornstahl's method is impracticable, the method suggested by J. T. Edsall and J. F. Foster (J. Am. Chem. Soc. 67, 617 (1945)) in which the length of the light path between the cylinder and observer is increased so as to reduce the intensity of the reflected light, may be used.

[26] A. Frey-Wyssling and E. Weber, Helv. Chim. Acta 24, 278 (1941).

[27] Y. Bjornstahl, Zeits. f. Physik 119, 3-4, 245 (1942).

of double refraction measurements made in a concentric cylinder apparatus and viscosity measurements made in a flow-tube viscometer are subject therefore to errors, which may be large for anomalous liquids. This point seems to have been overlooked by the majority of observers and suggests the use of a streaming double refraction apparatus with rotating outer cylinder and the inner cylinder freely suspended so that the whole forms a Couette viscometer.

It might well be mentioned here that the usual equation employed for obtaining the viscosity coefficient in a Couette viscometer is one deduced on the assumption that the liquid in the annular gap is Newtonian. If this equation is used for a non-Newtonian liquid the viscosity coefficient so obtained is an "apparent" coefficient equal to that of a Newtonian liquid producing the same deflection under identical conditions in the particular apparatus used.[28]

From the point of view of obtaining high values of the velocity gradient without turbulence there is theoretically an advantage in employing the outer cylinder as rotor although this advantage becomes less marked as the gap width is narrowed. However, from the practical point of view the difficulty experienced in the optical system with a narrow gap is greater than with a large gap.

Finally it must be remembered that the criteria for turbulence as given by Taylor hold only for Newtonian liquids and only when turbulence is due solely to centrifugal forces. Also it would appear that the ratio of cylinder length to gap width should be large. Any external disturbance, such as vibration, will tend to change the conditions and turbulence may therefore occur before the critical values given by Eq. (1) or the curves of Fig. 3. This was mentioned to the author in a private communication by Sir Geoffrey Taylor. Some such disturbance has in fact been detected by Buchheim, Stuart, and Menz[29] for even when allowance has been made for their mistake in applying Taylor's criterion incorrectly, turbulence sets in prematurely. Furthermore solutions which have long rod-shaped particles (i.e. which exhibit rod double refraction) may show turbulence earlier than Taylor's criteria would indicate. This is in accordance with the fact that the anomalous viscosity of such solutions will tend to change the value of the critical velocity (or the critical value of Reynolds number) for which turbulence occurs. Such a case has been observed by Andrade and Lewis[30] with ammonium oleate, the critical value of the velocity gradient being found to be about 30 percent lower than that given by Taylor's criterion for homogeneous liquids. This is probably the explanation of the anomalies observed by Hatschek and Jane.[31–33]

These facts indicate that it would be advantageous to have some physical indication of the onset of turbulence rather than having to rely upon calculations. This is possible if one adopts as a test for the onset of turbulence the fact that the relationship between viscous torque on the stationary cylinder and the speed of rotation of the moving cylinder instead of being linear becomes irregular. A brief discussion of this criterion as applied by Couette[34] and Mallock[35] is given by Taylor.[36]

A gradual departure from the linear law must, however, be carefully considered before it is taken as being due to turbulence for if the liquid under investigation is anomalous the torque may increase less quickly than the speed of rotation.[37]

VI. CONCLUSIONS

Tables I and II provide means by which an estimation may be made of the significance of the results of published work.

Theoretical and practical advantages in using the outer cylinder as rotor instead of the inner one are the following: (a) A larger gap can be used without the risk of turbulence and consequently greater ease is experienced in the optical measurements. (b) With a larger gap the temperature rise in the liquid is less. As the viscosity varies appreciably with temperature, the temperature of the liquid should be measured and not assumed to be that of the cooling bath. For this purpose it is advisable to incorporate a thermocouple in the concentric cylinder apparatus which may conveniently be built in the surface of the stationary cylinder. (c) The instrument can be used as a Couette viscometer so that double refraction and viscosity measurements can be made simultaneously.[38] (d) The onset of turbulence may be directly detected by irregularities in the torque-speed curve.

The onset of turbulence in pure liquids may not be accompanied by an abrupt change in the double refraction across the gap, although some change should occur consistent with the observed velocity distribution existing across the gap. A similar result holds for anomalous liquids, although it is probable that the velocity distribution across the gap under turbulent conditions differs slightly from that for a pure liquid.

---

[28] E. Hatschek, *Viscosity of Liquids* (G. Bell and Sons Ltd., London, 1928), p. 191. A. C. Merrington, *Viscometry* (E. Arnold & Company, London, 1949), p. 106.

[29] Buchheim, Stuart, and Menz, Zeits. f. Physik 112, (7–8), 407 (1939).

[30] E. N. da C. Andrade and J. W. Lewis, Kolloid. Zeits. 38, 260 (1926).

[31] E. Hatschek and R. S. Jane, Kolloid. Zeits. 38, 33 (1926).

[32] E. Hatschek, Kolloid. Zeits. 38, 259 (1926).

[33] W. Ostwald and R. Auerbach, Kolloid. Zeits. 38, 261 (1926).

[34] H. Couette, Ann. de. Chim. et de Phys. 6me sér., 21 (1890).

[35] A. Mallock, Phil. Trans. A187, 41 (1896).

[36] G. I. Taylor, Phil. Trans. A223, 289 (1923).

[37] E. Hatschek, Kolloid. Zeits. 13, 88 (1913). E. Hatschek and E. Humphrey, Proc. Phys. Soc. (London) 28, 274 (1916). E. Hatschek and R. S. Jane, Kolloid. Zeits. 40, 53 (1926). W. Ostwald, Kolloid. Zeits. 43, 210 (1927).

[38] For a discussion of the end effects in rotational viscometers see C. H. Lindsley and E. K. Fischer [J. App. Phys. 18, 988 (1947)]. The effect of the ends on the velocity distribution across the gap is treated by Taylor [Proc. Roy. Soc. A157, 572 (1936)].

In the light of the above discussion an instrument has been designed and is under construction; it is hoped that simultaneous measurements of viscosity and streaming double refraction will be possible.

VII. ACKNOWLEDGMENT

The author wishes to record the help and advice of Professor A. M. Taylor in whose laboratories research on streaming double refraction is being carried out.

Effective Stress and Effective Strain in Relation to Stress Theories of Plasticity*

G. N. White, Jr.** and D. C. Drucker†
Brown University, Providence, Rhode Island

(Received April 3, 1950)

The correlation of experimental data and the proper description of the state and history of deformation of work-hardening materials has received much attention. An often desired objective is the plotting of a variable involving stress alone, a so-called *effective stress*, against a variable depending on the instantaneous strain or the history of strain, *effective strain*. Quantities such as maximum or octahedral shearing stress and strain or better an integrated effective strain increment have been employed. Also, simple effective stress definitions have been given for anisotropic as well as isotropic metals by Dorn, by Jackson, Smith, Lankford, and by previous investigators.

It is shown that the stress-strain relations of a recent theory of plasticity provide a very convenient means of defining and studying effective stress and effective strain. The strong theoretical and practical limitations of such concepts are indicated and the calculation of specific plastic work is discussed for both isotropic and anisotropic metals.

One of the major objectives of the paper is to stimulate critical experimental examination of as yet unproved assumptions in the theory of plasticity and in the correlation of experimental data.

INTRODUCTION

THE relation between stress and strain for a work-hardening material is given for any one simple loading program (tension, shear, or any loading with constant ratios of the stress components) by a single curve. When the results obtained with different loading programs are to be correlated, or when the state and history of deformation for complex loading paths is to be described[1,2] it is again desirable to obtain a single curve. What is most often sought is a plot of some function of the stresses, termed the effective stress, against a function of the strain or strain history, termed effective strain. The related problem of equivalent proof stress has also been treated.[3]

The results of combined loading experiments have often been correlated by intuitively chosen stress and strain variables. Curves showing maximum shear stress *vs.* maximum shear strain are sometimes used. However, plots of octahedral shearing stress against octahedral shearing strain usually produce somewhat better correlation and are most often employed at present.

More recent definitions of effective stress $\bar{\sigma}$ and effective strain $\bar{\epsilon}$ have been developed in an entirely different manner to correlate not only simple loading programs but also those in which the directions and ratios of the principal stresses are not constant. Use is made of a stress-strain relation and a definition of effective strain based on the increment of specific plastic work, i.e. plastic work per unit volume, $dW^p = \bar{\sigma} d\bar{\epsilon}$. This approach has been employed for both isotropic and anisotropic materials.[4-7] If the effective stress $\bar{\sigma}$ in $dW^p = \bar{\sigma} d\bar{\epsilon}$ is taken to be the octahedral shearing stress, the increment of effective strain $d\bar{\epsilon}$ is found to be a function of the strain increments alone. For other definitions of $\bar{\sigma}$ it is by no means clear that the effective strain increment so defined will bear any simple relation to the measured or computed strain increments nor, in fact, have any direct physical significance.

The objective of this paper is to discuss these questions and to show how incremental stress-strain relations derived in terms of a criterion of yield or of further plastic deformation (termed a loading function) can be used in a systematic way to develop consistent and reasonable definitions of effective stress and effective strain. In this paper the loading function is chosen to be a function of the stress components only as is the current practice in experimental research. The choice of such a loading function, characteristic of the so-called "stress" theories of plasticity, leads directly to the definition of effective stress in terms of the stress com-

* The conclusions presented in this paper were obtained in the course of research conducted under contract N7onr-358 sponsored jointly by the ONR and the Bureau of Ships.

** Research Associate in Applied Mathematics, Brown University.
† Now Professor of Engineering, Brown University.

[1] W. Lode, Forsch. Gebiete Ingenieurw. No. 303 (1928).
[2] E. A. Davis, J. App. Mech., Trans. ASME 65, A187–196 (1943).
[3] J. Marin and R. L. Stanley, Am. Welding Soc. J., Supplement 745–805 (1940).
[4] C. Zener and J. H. Hollomon, Trans. A. S. M. 33, 163–214 (1944).
[5] Jackson, Smith, and Lankford, Am. Inst. Mining Met. Eng., Tech. Pub. No. 240, Class E, Metals Tech. August (1948).
[6] J. E. Dorn, J. App. Phys. 20, 15–20 (1949).
[7] J. C. Fisher, Trans. ASME. 71, 349–356 (1949).

ponents and, after some manipulation, to effective strain in terms of the history of strain. It should be kept in mind, however, that experimental evidence shows that the loading function also depends explicity on the strain (see reference 16). Moreover, effective stress must, in general, include stress history or the path of loading rather than final values of stress only. Therefore, the concepts of effective stress and strain as now employed and discussed in the following pages can have only a limited, although very useful, range of validity.

SYMBOLS

Note that repeated letter subscripts denote summation and any dummy letter may be used. For example, $\sigma_{ii} \equiv \sigma_{kk} = \sigma_{11}+\sigma_{22}+\sigma_{33} \equiv \sigma_x+\sigma_y+\sigma_z$ and $\sigma_{ij}d\epsilon_{ij} \equiv \sigma_{mn}d\epsilon_{mn} = \sigma_{11}d\epsilon_{11}+\sigma_{12}d\epsilon_{12}+\cdots+\sigma_{31}d\epsilon_{31}+\sigma_{33}d\epsilon_{33} = \sigma_x d\epsilon_x+\frac{1}{2}\tau_{xy}d\gamma_{xy}+\cdots+\frac{1}{2}\tau_{zx}d\gamma_{zx}+\sigma_z d\epsilon_z$.

c, a constant chosen to make effective stress reduce to a standard stress in some special case.

f, the loading function or criterion of plastic deformation; here taken as a function of the stresses alone.

$g=g(f)$ or $h(f)$, the part of the proportionality factor G in the stress-strain relations which depends on f alone.

m, an arbitrary constant in the proportionality factor G.

n, Poisson's ratio.

r, degree of f in the stresses.

$s_{ij} = \sigma_{ij}-(\sigma_{kk}/3)\delta_{ij}$, the stress deviator tensor $s_x = \frac{1}{3}(2\sigma_x-\sigma_y-\sigma_z)$, $s_{xy} = \tau_{xy}$, etc.

$t_{ij} = s_{ik}s_{kj}-\frac{2}{3}J_2\delta_{ij}$, the deviation of the square of s_{ij}.

A_{ijkl}, coefficient of the stresses in an anisotropic quadratic loading function.

A_{ij}, constants in the general anisotropic quadratic loading function.

B_{ij}, the inverse of A_{ij}, i.e. $B_{ij}A_{jk}=\delta_{ik}$.

C_{ijkl}, anisotropic elastic constants in the generalized Hooke's law.

A, B, C, D, E, F, constants in an orthotropic loading function.

E, Young's modulus.

G, the factor of proportionality in stress-strain relations: here a function of stress.

$H=H[(K_2)^{\frac{1}{2}}, (|K_3|)^{\frac{1}{3}}]$, a function of K_2, K_3 used to define an increment of effective strain.

$I(f)=H/df$, a function of f alone used to define an increment of effective strain.

$J_1=\sigma_{kk}$, the first invariant of the stress tensor σ_{ij} equal to $\sigma_x+\sigma_y+\sigma_z$.

$J_2=\frac{1}{2}s_{ij}s_{ij}$, the second order invariant of s_{ij}, the stress deviator or $\frac{1}{6}[(\sigma_1-\sigma_2)^2+(\sigma_2-\sigma_3)^2+(\sigma_3-\sigma_1)^2]$ equal to $\frac{1}{3}\tau_0^2$.

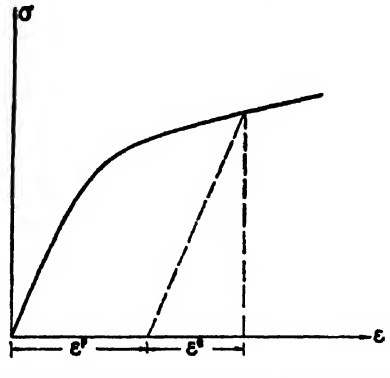


FIG. 1. Tension curve for a work hardening material.

$J_3=\frac{1}{3}s_{ij}s_{jk}s_{ki}$, the third order invariant of s_{ij}.

$K_2=de_{ij}^p de_{ij}^p$, second invariant of de_{ij}^p equal to $(de_x)^2+(de_y)^2+(de_z)^2+\frac{1}{2}(d\gamma_{xy})^2+\frac{1}{2}(d\gamma_{yz})^2+\frac{1}{2}(d\gamma_{zx})^2$.

$K_3=de_{ij}^p de_{jk}^p de_{ki}^p$, third invariant of de_{ij}^p.

$M_2=(\partial f/\partial\sigma_{ij})(\partial f/\partial\sigma_{ij})$, second invariant of $(\partial f/\partial\sigma_{ij})$, the gradient of f.

$M_3=(\partial f/\partial\sigma_{ij})(\partial f/\partial\sigma_{jk})(\partial f/\partial\sigma_{ki})$, third invariant of $(\partial f/\partial\sigma_{ij})$.

$R=(\frac{1}{2}M_2)^{\frac{1}{2}}/(\partial f/\partial\tau)$, the factor of difference between two stress-strain relations at a given value of f.

W^p, dW^p, specific plastic work and its increment.

δ_{ij}, the Kronecker delta equal to 0 for $i\neq j$, 1 for $i=j$.

$\bar{\epsilon}$, $\bar{\epsilon}^p$, effective strains, total and plastic.

$d\epsilon_{ij}=d\epsilon_{ij}^e+d\epsilon_{ij}^p$, increments of the total strain tensor ϵ_{ij} equal to the sum of the elastic and plastic increments with $d\epsilon_{11}=d\epsilon_x$, $d\epsilon_{12}=\frac{1}{2}d\gamma_{xy}$, $\cdots$, $d\epsilon_{33}=d\epsilon_z$.

ϵ_1, ϵ_2, ϵ_3, principal strains.

γ_0, octahedral shearing strain.

μ, ν, Lode's variables $\mu=(2\sigma_2-\sigma_1-\sigma_3/\sigma_1-\sigma_3)$, $\nu=(2d\epsilon_2-d\epsilon_1-d\epsilon_3/d\epsilon_1-d\epsilon_3)$.

$\bar{\sigma}$, effective stress.

σ_{ij}, stress tensor with components $\sigma_{11}=\sigma_x$, $\sigma_{12}=\tau_{xy}$, $\cdots$, $\sigma_{33}=\sigma_z$.

σ_1, σ_2, σ_3, principal stresses.

τ_0, octahedral shearing stress.

$\tau=\tau_{xy}$, shearing stress when $\tau_{yz}=\tau_{zx}=0$.

THE STRESS-STRAIN RELATION

The stress-strain relation to be used in this paper[8,9] is of the incremental or flow type which includes as special cases all explicit incremental laws now in use. A brief discussion of the development of the stress-strain relation will be given before the subject of effective stress and strain is treated. In this way the resulting definitions of effective stress and strain may be evaluated in the light of the assumptions made in deriving the stress-strain relation.

Figure 1 shows a typical tensile stress-strain curve for a work-hardening material. It is assumed that time and temperature effects are small enough to be disregarded. The stress-strain curve, which has positive slope, for the entire range considered represents then a series of states of stable static equilibrium. As no simple correlation between stress and strain exists for complex loading paths, the meaning of work-hardening for the general case requires consideration.

Work-hardening may be defined in the following manner: If for all additional sets of stresses which are slowly applied to and then removed from a block of stressed material in a given state of deformation by an external agency, the material remains in equilibrium and (a) positive work is done *by the external agency* during the application of the stresses and (b) the net work performed by it over the cycle of its application and removal is positive (or zero if no plastic deformation occurs), then the material is said to be work hardening. In other words, no work can be extracted in a cycle from the system including the block of material and the forces acting upon it.

The stress-strain relations which are derived in con-

[8] W. Prager, J. App. Phys. 20, 235–241 (1949).
[9] D. C. Drucker, J. Coll. Sci., Rheology Issue 4, 299–311 (1949). D. C. Drucker, Quart. App. Math. 7, 411–418 (Jan. 1950); D. C. Drucker, Trans. A. S. M. E. 71, 587–592 (1949).

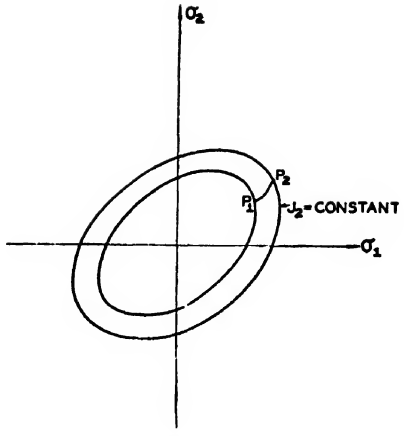


FIG. 2. A level curve of $f=J_2=\frac{1}{3}\tau_0^2$.

formity with the work hardening conditions are based on two assumptions:

(1) The relation between the increments of stress and strain is linear (all explicit stress-strain relations proposed so far adopt this assumption).

(2) A yield criterion or loading function f exists.

At any given state of plastic deformation a function of the stress components, $f(\sigma_{ij})=f(\sigma_x, \sigma_y, \cdots, \tau_{yz})$, exists such that further plastic deformation takes place only if f increases. In the general theory the value of the loading function f and also the form of f will depend on the plastic strain and plastic strain history.

In order to write the stress-strain relations compactly and to express basic ideas clearly, tensor notation will be employed. For example, the state of stress is represented by the single symbol σ_{ij} which stands for the stress tensor with the components $\sigma_{11}=\sigma_x$, $\sigma_{12}=\tau_{xy}$, $\cdots$, $\sigma_{33}=\sigma_z$. Moreover use of the tensor convention of summing over repeated subscripts, results in simplification of formulae and in case of manipulation, e.g. $a_ib_i=a_1b_1+a_2b_2+a_3b_3$, and $a_{ii}=a_{11}+a_{22}+a_{33}$.

The work-hardening conditions and the two assumptions lead to an expression for the plastic strain increments $d\epsilon_{ij}^p$ in terms of the gradient of the loading function $\partial f/\partial\sigma_{ij}$, a proportionality factor G, and the stress increments $d\sigma_{kl}$. In tensor notation the relation is

$$d\epsilon_{ij}^p=G\frac{\partial f}{\partial\sigma_{ij}}\left(\frac{\partial f}{\partial\sigma_{kl}}d\sigma_{kl}\right). \tag{1}$$

The factor G is in general a function of the stress, strain, and strain history. The positive quantity $(\partial f/\partial\sigma_{kl})d\sigma_{kl}$ is the increase in the loading function resulting from stress changes only; for example $d\tau_0$, if τ_0, the octahedral shearing stress, is the criterion of loading. The total increment of strain $d\epsilon_{ij}$ is obtained by adding $d\epsilon_{ij}^e$, the elastic increment, to the plastic increment $d\epsilon_{ij}^p$. A general anisotropic form of Hooke's law, $d\epsilon_{ij}^e=C_{ijkl}d\sigma_{kl}$, or the usual isotropic one may be used.

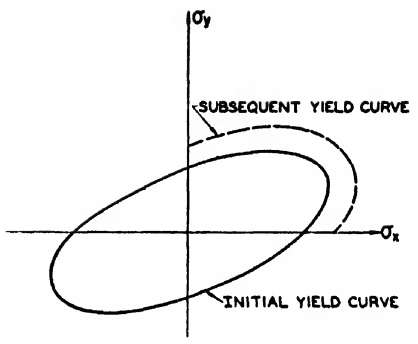


FIG. 3. Intrinsic anisotropy $f=f(\sigma)$, the flow stress in simple tension in the x-direction is not the same as for the y-direction.

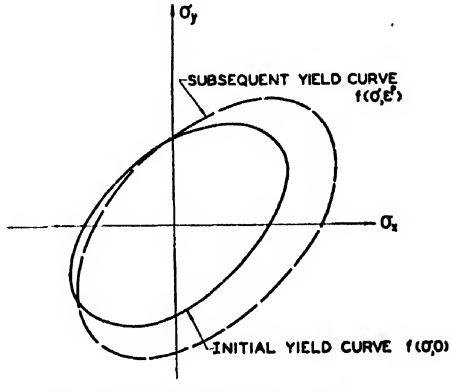


FIG. 4. Bauschinger effect—$f=f(\sigma, \epsilon^p)$.

The two resulting expressions are respectively

$$d\epsilon_{ij}=d\epsilon_{ij}^e+d\epsilon_{ij}^p=C_{ijkl}d\sigma_{kl}+G\frac{\partial f}{\partial\sigma_{ij}}\left(\frac{\partial f}{\partial\sigma_{kl}}d\sigma_{kl}\right) \tag{2}$$

and

$$d\epsilon_{ij}=\frac{(1+n)}{E}d\sigma_{ij}-\frac{n}{E}\delta_{ij}d\sigma_{kk}+G\frac{\partial f}{\partial\sigma_{ij}}\left(\frac{\partial f}{\partial\sigma_{kl}}d\sigma_{kl}\right), \tag{3}$$

where n is Poisson's ratio, E is Young's modulus, and δ_{ij} is the Kronecker delta, which is zero when $i\neq j$, unity when $i=j$. If the simplest loading function, the Mises' yield condition $f=J_2$ (shear strain energy or square of the octahedral shearing stress to within a multiplicative constant) is used, Eq. (3) in engineering notation is

$$\left.\begin{aligned} d\epsilon_x&=\frac{1}{E}[d\sigma_x-n(d\sigma_y+d\sigma_z)]\\ &\qquad+g(\tau_0)[\sigma_x-\tfrac{1}{2}(\sigma_y+\sigma_z)]d\tau_0\\ 2d\epsilon_{xy}=d\gamma_{xy}&=\frac{2(1+n)}{E}d\tau_{xy}+3g(\tau_0)\tau_{xy}d\tau_0 \end{aligned}\right\}. \tag{4}$$

These forms have been used by Nadai[10] and later by Dorn.[11]

Figure 2, a plot of Mises' loading function J_2=constant for a two-dimensional stress system, reveals many of the features of the more general loading functions. According to the postulates for the stress-strain relations, if the stresses change so as to move the point P_1 representing them outside of the ellipse J_2=constant on which the point is located, plastic deformation will take place and a new ellipse is established through the new point P_2. If the stresses change so as to move the point inside or on the ellipse, only elastic deformation can occur. In general, at any stage of plastic deformation, the loading function $f(\sigma_{ij})$ may be regarded as a surface in a multi-dimensional space with coordinates represented by all those stress components σ_{ij} which are not kept zero during the test. For a given state of stress at a point in the material, it is a surface which bounds all those states of stress which can be reached without additional plastic deformation taking place. As the increment of plastic work is necessarily positive, the surface f must be convex with respect to the origin as illustrated by Fig. 2.

An advantage of the general stress-strain relations Eqs. (2) and (3) is that anisotropic materials may be treated almost as easily as isotropic. Two types of anisotropy are (a) an initial anisotropy which is not altered appreciably by subsequent deformation, the loading function depending explicitly upon the stress components alone $f=f(\sigma_{ij})$, and (b) anisotropy produced by the plastic deformation being considered, the loading function depending explicitly on the components of plastic strain as well as stress, e.g. $f=f(\sigma_{ij}, \epsilon_{ij}^p)$. The first type of anisotropic loading function may be useful for a material which has been strongly cold worked prior to testing; the second for a material which is initially isotropic and exhibits a marked Bauschinger effect. Two-dimensional illustrations for particular examples of the two types of anisotropy are given in Figs. 3 and 4.

As f is here taken to depend explicitly on stress alone, the most general form of isotropy considered is that produced by using a loading function which is a function of the stress invariants J_1, J_2, and J_3. The octahedral shearing stress, $\tau_0=(\frac{2}{3}J_2)^{\frac{1}{2}}$, is a special case where $J_1=\sigma_{kk}$, the sum of the principal stresses, and J_3, which is a weighting factor for shearing stresses, are assumed to have no influence. The stress-strain relation (3) for $f=J_2=\frac{3}{2}\tau_0^2$ and $G=G(f)$ has been given in engineering notation, Eq. (4). In the simple tensor form, the plastic strain increment is

$$d\epsilon_{ij}^p=G(J_2)s_{ij}dJ_2=h(\tau_0)(\sigma_{ij}-(\sigma_{kk}/3)\delta_{ij})d\tau_0 \quad (5)$$

as s_{ij}, the stress deviator, is $\sigma_{ij}-(\sigma_{kk}/3)\delta_{ij}$. G and h are functions of J_2, or equivalently, of τ_0. A material obeying the relation (5) which uses J_2 as the loading function is called isotropic work hardening as it exhibits no directional properties when unloaded. There is no way of distinguishing the type of loading which produced the work hardening. All tensile specimens cut from the strained material will have the same yield point no matter what their orientation. Also, if the material is loaded a given amount in tension in any direction the extent of yielding or the value of f will be the same as for any other direction. Although the relation (5) is isotropic in the sense described, writing all the terms shows that when the material is under load (5) has the same form as a general anisotropic Hooke's law:

$$d\epsilon_x=G_1d\sigma_x+G_2d\sigma_y+G_3d\sigma_z+G_4d\tau_{xy}+G_5d\tau_{yz}+G_6d\tau_{zx}, \quad (6)$$

where the G's depend on stress alone. However, the anisotropy is produced entirely by the existing state of stress; it does not subsist after unloading.

Experimental evidence clearly indicates the need for more general stress-strain relations. The next extension is to retain the assumption of isotropic work hardening but include the third invariant of the stresses J_3 as well as J_2. The form of the relation (1) becomes then

$$d\epsilon_{ij}^p=G\left\{\frac{\partial f}{\partial J_2}s_{ij}+\frac{\partial f}{\partial J_3}t_{ij}\right\}df, \quad (7)$$

where t_{ij} is the deviation of the square of s_{ij}, $s_{ik}s_{kj}-\frac{2}{3}J_2\delta_{ij}$, and G is a function of stress.

If an intrinsically anisotropic stress-strain relation is needed, the loading function can no longer be a function of the stress invariants J_2 and J_3 but must depend directly on stress components referred to fixed axes. A simple convenient choice is the anisotropic quadratic form with constant coefficients $f=A_{ijkl}\sigma_{ij}\sigma_{kl}$ which reduces to J_2 in the isotropic case if the sum of the principal stresses is without influence. The stress-strain relation then obtained from Eq. (1) is

$$d\epsilon_{ij}^p=2GA_{ijkl}\sigma_{kl}df \quad (8)$$

where G is a function of stress. In expanded form using $\tau_{ij}=\tau_{ji}$

$$\left.\begin{aligned} d\epsilon_x^p&=2Gdf\{A_{1111}\sigma_x+2A_{1112}\tau_{xy}+A_{1122}\sigma_y \\ &\quad +2A_{1123}\tau_{yz}+A_{1133}\sigma_z+2A_{1131}\tau_{zx}\} \\ 2d\epsilon_{xy}^p=d\gamma_{xy}^p&=4Gdf\{A_{1211}\sigma_x+2A_{1212}\tau_{xy}+A_{1222}\sigma_y \\ &\quad +2A_{1223}\tau_{yz}+A_{1233}\sigma_z+2A_{1213}\tau_{zx}\} \end{aligned}\right\} \quad (9)$$

where the A's are just numbers. For a material in which there exist planes or axes with respect to which the anisotropy is symmetric, the general quadratic loading function and the corresponding stress-strain relation may be simplified. If for example the material is orthotropic and the principal axes of stress coincide with the lines of intersection of the three mutually orthogonal planes of symmetry taken as coordinate planes, the loading function is $f=a_{ij}\sigma_i\sigma_j$ $(i,j=1, 2, 3)$ where the stresses σ_i are principal stresses $\sigma_1=\sigma_x$ etc.

[10] A. Nadai, J. App. Phys. 8, 205–212 (1937).
[11] Cunningham, Thompson, and Dorn, Am. Soc. Test. Mat. 47, 546–553 (1947).

A typical plastic strain increment is then

$$d\epsilon_{ij}^{p} = 2Gdf(a_{11}\sigma_x + a_{12}\sigma_y + a_{13}\sigma_z). \quad (10)$$

Eqs. (9) and (10) are stress-strain relations used by Dorn, by Jackson, Smith, and Lankford, and by Fisher.

More elaborate forms of f may easily be written as required by experimental data.

EFFECTIVE STRESS AND EFFECTIVE STRAIN

Effective stress $\bar{\sigma}$ is by definition a quantity which determines whether or not further plastic deformation takes place. Since it must be defined consistently with the stress σ in a simple tension test of a work hardening material, Fig. 1, $\bar{\sigma}$ must be a positive increasing function of the stresses during plastic deformation. For stress theories of plasticity, $f=f(\sigma_{ij})$, the loading function f also by definition determines whether additional plastic deformation will occur. It too is a positive increasing function as long as plastic deformation continues, and if unloading occurs, plastic deformation is not resumed until the highest previous value of f is exceeded. For stress theories there is then no real choice in the definition of the effective stress $\bar{\sigma}$. The loading function f is the effective stress to some power and to within an arbitrary constant; for example $\tau_0=(\frac{2}{3}J_2)^{\frac{1}{2}}$ or $\bar{\sigma}=(3J_2)^{\frac{1}{2}}$.[10,11]

On the other hand the definition of an effective strain $\bar{\epsilon}$ related to the effective stress is not made so easily. The method now used is to arbitrarily define the permanent effective strain $\bar{\epsilon}^{p}$ by

$$dW^{p} = \bar{\sigma}d\bar{\epsilon}^{p}, \quad (11)$$

where dW^{p} is the increment of specific plastic work of deformation.[4-7] This definition is based on the belief that the plastic work is a fundamental measure of the amount of plastic deformation.

The definition (11) of $\bar{\epsilon}$ in terms of plastic rather than total work is made to avoid considering increases of total work in which plastic deformation plays no part. Such a situation arises when the loading path is one for which $f=$constant. In this case the plastic increments of strain (1) are zero whereas the elastic strain increments are not.

A second basic method is the use of the stress-strain relations, Eq. (1), to find an integral of the strain increments which is exactly or quite closely a function of the loading function f or effective stress $\bar{\sigma}$ only. The integral of the strain increments is then defined to be the effective strain.

The second method is an attempt to define an effective strain which has a direct physical meaning in terms of measured strain and strain history. The first method does not in general lead to an effective strain which is related to strain as ordinarily defined. The effective strain increment cannot be computed from the strains and their history only, as may be seen when Eq. (11) is expressed fully in terms of the stresses and strain increments by means of Eq. (1).

$$dW^{p} = \bar{\sigma}d\bar{\epsilon}^{p} = \sigma_{ij}d\epsilon_{ij}^{p} = \sigma_{ij}G\frac{\partial f}{\partial \sigma_{ij}}\left(\frac{\partial f}{\partial \sigma_{kl}}d\sigma_{kl}\right). \quad (12)$$

It should be kept in mind that the concept of effective stress $\bar{\sigma}$ and consequently of effective strain $\bar{\epsilon}^{p}$ as presented here is only valid for materials with negligible Bauschinger and similar effects, or the more usual case of paths of loading which do not reveal such effects directly. Figure 4 shows that to account for Bauschinger effect the loading function must depend on the plastic strains as well as on stress. An effective stress which depends on stress only is then no longer possible as consideration of alternating tension and compression clearly shows.

In the following sections the expressions for $\bar{\sigma}$ and $\bar{\epsilon}^{p}$ will be given explicit forms in accordance with certain other features of the loading functions and stress-strain relations dictated by experimental results.

ISOTROPIC THEORIES

The simplest form of stress-strain relation for a material which work hardens isotropically, Eq. (5), will be used first as an example of both methods for defining the effective strain $\bar{\epsilon}^{p}$ in terms of the effective stress $\bar{\sigma}$. In this case $\bar{\sigma}$ may be defined as $(f)^{\frac{1}{2}}$ or $(J_2)^{\frac{1}{2}}$ multiplied by any constant. Choosing $\bar{\sigma}=(3J_2)^{\frac{1}{2}}$ makes $\bar{\sigma}=\sigma_x$ for uniaxial tension in the x-direction. The resulting $\bar{\epsilon}^{p}$ is then the same as Dorn's.[11]

Using the first method and noting that $(\partial J_2/\partial\sigma_{ij})=s_{ij}$ and $(d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}}=G(2J_2)^{\frac{1}{2}}dJ_2$ gives

$$\begin{aligned}\bar{\sigma}d\bar{\epsilon}^{p} &= dW^{p} = \sigma_{ij}d\epsilon_{ij}^{p} = \sigma_{ij}G(J_2)(\partial J_2/\partial\sigma_{ij})dJ_2 \\ &= (3J_2)^{\frac{1}{2}}(\tfrac{2}{3}d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}} = \bar{\sigma}(\tfrac{2}{3}d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}}. \end{aligned} \quad (13)$$

Consequently the effective strain increment is

$$\begin{aligned}d\bar{\epsilon}^{p} = (\tfrac{2}{3}d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}} \equiv (\tfrac{2}{3})^{\frac{1}{2}}[&(d\epsilon_x^{p})^2+(d\epsilon_y^{p})^2 \\ &+(d\epsilon_z^{p})^2+\tfrac{1}{2}(d\gamma_{xy}^{p})^2+\tfrac{1}{2}(d\gamma_{yz}^{p})^2+\tfrac{1}{2}(d\gamma_{zx}^{p})^2]^{\frac{1}{2}}, \end{aligned} \quad (14)$$

as $d\gamma_{xy}=2d\epsilon_{xy}$, etc. and $d\epsilon_{xy}=d\epsilon_{yz}$.

On the other hand, the second method uses the stress-strain relation Eq. (5) directly to form an increment of effective strain. The simplest combination of the plastic strain increments which is positive increasing and has the correct dimension for an effective strain increment is $(d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}}$ so that with $(\frac{2}{3})^{\frac{1}{2}}$ inserted arbitrarily to make $\bar{\epsilon}^{p}=\epsilon_z^{p}$ for uniaxial tension in the x-direction

$$d\bar{\epsilon}^{p} = (\tfrac{2}{3})^{\frac{1}{2}}(d\epsilon_{ij}^{p}d\epsilon_{ij}^{p})^{\frac{1}{2}} = (\tfrac{2}{3})^{\frac{1}{2}}G(J_2)(2J_2)^{\frac{1}{2}}dJ_2 = G^{*}(\bar{\sigma})d\bar{\sigma} \quad (15)$$

which shows that the integrated effective strain $\bar{\epsilon}^{p}$ is a function of the effective stress $\bar{\sigma}$. Once the factor $G(J_2)$ has been found for a given material by matching the stress-strain relation, Eq. (5), to an experimental curve, the formula for $\bar{\epsilon}^{p}$ in terms of $\bar{\sigma}$ is completely determined. In this approach also the constant $\sqrt{3}$ in $\bar{\sigma}$ may be inserted if desired to make $\bar{\sigma}=\sigma_x$ for simple tension in the x-direction.

Before the definition of $\bar{\epsilon}^p$ is extended by means of each of the two methods to materials requiring more complicated loading functions, two features which account for the simplicity of the preceding result should be noted. The first is that the loading function $f=J_2=\frac{1}{3}\tau_0^2$ is homogeneous (quadratic) in the stresses; the second, that G is taken to depend on the loading function only. If these two properties are assumed to hold for other stress loading functions of degree r in the stresses and for the corresponding stress-strain relations, the expression for the specific plastic work is found to be a function of f

$$dW^p=\sigma_{ij}G(f)\frac{\partial f}{\partial\sigma_{ij}}df=rfG(f)df, \tag{16}$$

which, if true, would mean that when plastic deformation takes place the plastic work is independent of the path of loading.

Some evidence that f is a homogeneous function of the stresses exists in the results of Fraenkel's tube tests.[12] It was found that for a tube loaded in tension and internal pressure applied in constant ratio, $\mu-\nu$, the difference between Lode's variables, is quite constant for octahedral shear strains larger than 0.02 percent (see discussion[13]). Since ν is given by

$$\nu=\frac{2d\epsilon_2^p-d\epsilon_1^p-d\epsilon_3^p}{d\epsilon_1^p-d\epsilon_3^p}$$

$$=\frac{2(\partial f/\partial\sigma)_2-(\partial f/\partial\sigma)_1-(\partial f/\partial\sigma)_3}{(\partial f/\partial\sigma)_1-(\partial f/\partial\sigma)_3}, \tag{17}$$

homogeneity means constancy of ν as found in the experiment.

A lack of dependence of the dissipated work on the path of loading is a rather surprising result for a nonconservative system. However, Schmidt[14] has shown that, for tubes tested under tension and torsion applied in ratio, equally satisfactory correlation is obtained by plotting $f=\tau_0$ *vs.* work and $f=\tau_0$ *vs.* γ_0. On the other hand, a study of the results of Taylor and Quinney[15] indicates that, in their tests at least, the work dissipated is strongly path dependent and does not depend on the loading function alone. This is one of the most important questions still unanswered and much experimental work remains to be done.

When the effective stress $\bar{\sigma}$ and strain $\bar{\epsilon}^p$ of the simple J_2 or τ_0 isotropic theory do not give satisfactory correlation of experimental data,[16] definitions of $\bar{\sigma}$ and $d\bar{\epsilon}^p$ which depend on a stress-strain relation including the third stress invariant J_3, Eq. (7), should be tried. The necessity for including J_3 to fit experimental results is apparent from the fact that the J_2 theory gives straight μ *vs.* ν curves, i.e. $\mu=\nu$, whereas experiments show ν to be numerically less than μ except at special points. For isotropic loading functions this means that f depends on both J_2 and J_3. A possible inference is that the effective strain include not only the second strain increment invariant $K_2=d\epsilon_{ij}^p d\epsilon_{ij}^p$, as in the J_2 theory, but also the third invariant $K_3=d\epsilon_{ij}^p d\epsilon_{jk}^p d\epsilon_{ki}^p$. However, it will be shown that $(K_2)^{\frac{1}{2}}$ alone provides a satisfactory effective strain increment for many correlations of experimental data.

In the more general isotropic theory $f=f(J_2, J_3)$ where f is homogeneous of degree r in the stresses. Then $\bar{\sigma}=\bar{\sigma}(J_2, J_3)=cf^{1/r}$ where c is a constant which may be adjusted arbitrarily, as before, to make $\bar{\sigma}=\sigma_x$ in the case of simple tension. The first definition of the effective strain $\bar{\epsilon}^p$ by means of $dW^p=\bar{\sigma}d\bar{\epsilon}^p$ is useful but one without direct physical meaning as it is a function of stress as well as strain and cannot be computed from the measurement of strain and strain history along. This approach is equivalent to plotting plastic work against effective stress.[14] On the other hand, the second method for defining the effective strain using the stress-strain relations does not necessarily load to an $\bar{\epsilon}^p$ which is precisely dependent on the loading function f alone. Hence one may have to choose an $\bar{\epsilon}^p$ which is very closely related to, but is not truly a function of, $\bar{\sigma}$ alone. It will, in fact, be shown that for $f(J_2, J_3)$ the effective strain as defined for the simpler J_2 theory is satisfactory for experimental purposes providing, of course, that a J_2, J_3 theory is adequate for the data.

Consider the definition of $d\bar{\epsilon}^p$ by

$$\left(\frac{3}{2}\right)^{\frac{1}{2}}d\bar{\epsilon}^p=(d\epsilon_{ij}^p d\epsilon_{ij}^p)^{\frac{1}{2}}\equiv(K_2)^{\frac{1}{2}}$$

$$=G(f)\left(\frac{\partial f}{\partial\sigma_{ij}}\frac{\partial f}{\partial\sigma_{ij}}\right)^{\frac{1}{2}}df$$

$$=G(f)(M_2)^{\frac{1}{2}}df \tag{18}$$

where M_2 is the second invariant of $\partial f/\partial\sigma_{ij}$, the gradient of f. For the isotropic loading functions under consideration Eq. (18) may be written

$$\left(\frac{3}{2}\right)^{\frac{1}{2}}d\bar{\epsilon}^p=G(f)\left[2J_2\left(\frac{\partial f}{\partial J_2}\right)^2+6J_3\frac{\partial f}{\partial J_2}\frac{\partial f}{\partial J_3}\right.$$

$$\left.+\frac{2}{3}J_2^2\left(\frac{\partial f}{\partial J_3}\right)^2\right]df. \tag{19}$$

It is apparent from (19) that in general this choice of $d\bar{\epsilon}^p$ is not a function of f and df alone and cannot be integrated to give $\bar{\epsilon}^p$ as a function of f alone. Since this isotropic theory includes the effect of J_3, a definition of

[12] S. J. Fraenkel, J. App. Mech., Trans. ASME **70**, 193–200 (1948).
[13] D. C. Drucker and W. Prager, discussion of reference 12, J. App. Mech., Trans. ASME **71**, 101–102 (1949).
[14] R. Schmidt, Ingenieur Archiv. **3**, 215–235 (1932).
[15] G. I. Taylor and H. Quinney, Phil. Trans. Roy. Soc. **A230**, 323–362 (1931).
[16] D. C. Drucker, J. App. Mech., Trans. ASME **71**, 349–357 (1949).

$d\bar{\epsilon}^p$ may be tried which includes the third invariant K_3 as well as K_2, e.g.,

$$d\bar{\epsilon}^p = H([K_2]^{\frac{1}{2}}, [|K_3|]^{\frac{1}{3}}). \tag{20}$$

While neither K_2 nor K_3 are functions of f and df alone, the function $H([K_2]^{\frac{1}{2}}, [|K_3|]^{\frac{1}{3}})$ may be equivalent to a function I of f times df

$$d\bar{\epsilon}^p = H([K_2]^{\frac{1}{2}}, [|K_3|]^{\frac{1}{3}}) = I(f)df \tag{21}$$

and therefore may be integrated to give $\bar{\epsilon}^p = \bar{\epsilon}^p(f) = \bar{\epsilon}^p(\bar{\sigma})$. Although a relation of this type between H and I must always exist, $I(f)$ is given by an infinite series in J_2 and J_3, except in degenerate cases such as $f = J_2$. Definitions of $d\bar{\epsilon}^p$ of the type of Eq. (21) are therefore of little practical value. A restricted extension of the second method for defining $\bar{\epsilon}^p$ for $f = f(J_2, J_3)$ and $G = G(f)$

$$\bar{\epsilon}^p = \bar{\epsilon}^p\left(\int (K_2)^{\frac{1}{2}}dt, \quad \int (|K_3|)^{\frac{1}{3}}dt\right) = \bar{\epsilon}^p(f) \tag{22}$$

has the same defects as the definition of Eq. (21); a relation between the two integrals and f in (22) does exist but cannot generally be found in a simple, useful form.

These general procedures fail to give a useful $\bar{\epsilon}^p$, but a satisfactory $\bar{\epsilon}^p$ can be developed by means of the stress-strain relations Eq. (1) as the following example shows. Suppose that in Eq. (1) the function G is modified slightly to

$$G = g(f)/(M_2)^{\frac{1}{2}} \tag{23}$$

instead of $G = G(f)$ (as yet no conclusive experimental evidence exists for or against such a choice). The plastic strain increments are then

$$d\epsilon_{ij}^p = \frac{g(f)}{(M_2)^{\frac{1}{2}}} \frac{\partial f}{\partial \sigma_{ij}} df. \tag{24}$$

Squaring each term and adding gives

$$d\epsilon_{ij}^p d\epsilon_{ij}^p = g^2(f)(df)^2, \tag{25}$$

so that $(d\epsilon_{ij}^p d\epsilon_{ij}^p)^{\frac{1}{2}} = g(f)df$ when integrated will give a function of f alone. Thus any constant times $(d\epsilon_{ij}^p d\epsilon_{ij}^p)^{\frac{1}{2}}$ is an effective strain increment; its integral $\bar{\epsilon}^p$ depends on f or the effective stress $\bar{\sigma}$ alone. The function G is easily modified to produce more complicated definitions of the effective strain in which both K_2 and K_3 appear. If G is

$$G = h(f)/(M_2)^{\frac{1}{2}} + m(|M_3|)^{\frac{1}{3}} \tag{26}$$

where

$$M_2 = \frac{\partial f}{\partial \sigma_{ij}} \frac{\partial f}{\partial \sigma_{ij}}, \quad M_3 = \frac{\partial f}{\partial \sigma_{ij}} \frac{\partial f}{\partial \sigma_{jk}} \frac{\partial f}{\partial \sigma_{ki}},$$

and m is a constant, then the effective strain may be taken as any constant times

$$d\bar{\epsilon}^p = h(f)df = (K_2)^{\frac{1}{2}} + m(|K_3|)^{\frac{1}{3}}, \tag{27}$$

which again may be integrated directly. Still more general combinations, of course, may be derived.

These altered forms of the stress-strain relation are basically different from the previous ones for which $G = G(f)$. The work dissipated W^p is no longer dependent on the final state of stress or f alone. At present there is no definite indication as to which type of assumption is closer to physical reality. Experiment alone must decide, in each individual case, which stress-strain relation is best. As mentioned previously, the process of plastic deformation is irreversible so that even limited path independence $W^p = W^p(f)$ would be quite startling.

Fortunately, as the next section demonstrates, comparison of the results obtained with the various forms of G shows that the difference between $\bar{\epsilon}^p$ *vs.* $\bar{\sigma}$ curves in which $d\bar{\epsilon}^p$ is defined by means of the different factors is small. *Therefore $d\bar{\epsilon}^p$ can be chosen as $(\frac{2}{3})^{\frac{1}{2}}(d\epsilon_{ij}^p d\epsilon_{ij}^p)^{\frac{1}{2}}$ for materials which work harden isotropically in the range considered.*

THE DEPENDENCE OF THE EFFECTIVE STRAIN $\bar{\epsilon}^p$ ON THE FACTOR G

It having been demonstrated that the effective strain $\bar{\epsilon}^p = \int (\frac{2}{3})^{\frac{1}{2}}(d\epsilon_{ij}^p d\epsilon_{ij}^p)^{\frac{1}{2}}$ is a function of f or $\bar{\sigma}$ when the factor G in the stress-strain relation is given by $g(f)/(M_2)^{\frac{1}{2}}$ whereas $d\bar{\epsilon}^p$ is not in general a function of f along when $G = G(f)$, it is desirable to find a way of estimating the difference in $\bar{\epsilon}^p$ according to the two different types of G.

The two stress-strain relations may be written

$$d\epsilon_{ij}^{p(A)} = G(f) \frac{\partial f}{\partial \sigma_{ij}} df \tag{28}$$

$$d\epsilon_{ij}^{p(B)} = \frac{g(f)}{(M_2)^{\frac{1}{2}}} \frac{\partial f}{\partial \sigma_{ij}} df. \tag{29}$$

If the two relations are each matched to a given experimental shear stress *vs.* shear strain curve, τ *vs.* γ, it is readily found that, for an increment of f at any value of f, the increments of plastic strain for any loading path are related to each other quite simply by

$$d\epsilon_{ij}^{p(A)} = \frac{(\frac{1}{2}M_2)^{\frac{1}{2}}}{(\partial f/\partial \tau)} d\epsilon_{ij}^{p(B)} = R d\epsilon_{ij}^{p(B)}. \tag{30}$$

where $R = (\frac{1}{2}M_2)^{\frac{1}{2}}/(\partial f/\partial \tau)$. The maximum difference between the two strains can be computed for a given loading function and given types of loading by finding the deviation of R from unity.

For the extreme case of the particular loading function $f = J_2^3 - (9/4)J_3^2$ in which J_3, the shearing stress weighting factor, is much more prominent than ordinarily demanded by experimental data, the range of variation of the factor of difference R is only $0.935 \leq R \leq 1$. Figure 5 shows the maximum difference in the

tension curves for the two laws after they have been matched in shear. The curve A is obtained by multiplying the strain value for the B curve by $R=0.935$. If instead of a tension curve, the effective stress-strain curve $\bar{\sigma}$ *vs.* $\bar{\epsilon}^p$ is desired, the same method is applied since $d\bar{\epsilon}^{p(A)}=(\frac{2}{3})^{\frac{1}{2}}(d\epsilon_{ij}{}^{p(A)}d\epsilon_{ij}{}^{p(A)})^{\frac{1}{2}}=Rd\bar{\epsilon}^{p(B)}$ according to the second method of defining effective strain.

The difference between the plots of $\bar{\sigma}$ *vs.* $\bar{\epsilon}^p$ for any other loading path, no matter how complicated, is less than for simple tension. Therefore, if effective strain $\bar{\epsilon}^p$ is to be plotted as a function of $\bar{\sigma}$, one may use $\bar{\epsilon}^p=(\frac{2}{3})^{\frac{1}{2}}\int(d\epsilon_{ij}{}^p d\epsilon_{ij}{}^p)^{\frac{1}{2}}$ with the assurance that even for extreme cases a small variation only occurs between a plot for which Eq. (28) applies and one for which Eq. (29) is correct. The effective strain, so defined, may be used to calculate the plastic work done by $W^p=\int\bar{\sigma}d\bar{\epsilon}^p$ within close limits providing, of course, that the assumption of isotropic work hardening is sufficiently good for the path of loading studied.

It is interesting to note that in the customary tests on thin walled tubes in which all stresses are increased in ratio, the simple $\bar{\epsilon}^p$ defined above is a constant times the octahedral shearing strain. If the material of the tube is reasonably isotropic at the start, all such tests, no matter what the ratios chosen, will give very closely

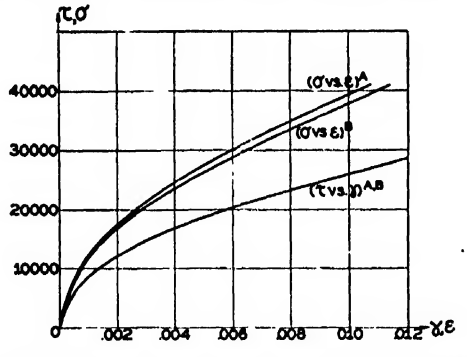


FIG. 5. Predicted σ *vs.* ϵ curves for A and B types of stress-strain relations based on a τ *vs.* γ curve of a severe test with the effect of J_3 overly prominent.

a single curve when the coordinates are chosen as $\bar{\sigma}(J_2, J_3)$ and $\bar{\epsilon}^p$ as the octahedral shearing strain.[16]

The expressions for effective stress $\bar{\sigma}=\bar{\sigma}(f)$ and effective strain $\bar{\epsilon}^p$ developed so far have been for isotropic work hardening and will not correlate experimental information from initially anisotropic materials. The general definitions of $\bar{\sigma}$ and $\bar{\epsilon}^p$, of course, also apply to the anisotropic case, but there the proper anisotropic stress-strain relations must be used.

ANISOTROPIC THEORY

As stated, the "stress" theories of plasticity $f=f(\sigma_{ij})$ with which this paper is concerned can only treat anisotropic materials which are free from Bauschinger effects for the loading being considered. The anisotropy is regarded as intrinsic in the material and is manifested in the loading function by asymmetry.

The work of Dorn,[6] of Fisher,[7] of Jackson, Smith and Lankford,[5] and of Hill[17] indicates that for many anisotropic materials, a loading function quadratic in the stresses is almost as satisfactory as J_2 is for the isotropic case. For such a loading function the stress-strain relation is given by Eq. (8)

$$d\epsilon_{ij}{}^p=2GA_{ijkl}\sigma_{kl}df. \tag{8}$$

As an example, the case of orthotropic symmetry in a material for which hydrostatic pressure does not influence plastic deformation and G is a function of f will be considered. The loading function f can be written as

$$\begin{aligned}2f=&A(\sigma_x-\sigma_y)^2+B(\sigma_y-\sigma_z)^2+C(\sigma_z-\sigma_x)^2+3D\tau_{xy}{}^2\\&+3D\tau_{yz}{}^2+3E\tau_{yz}{}^2+3E\tau_{zy}{}^2+3F\tau_{zx}{}^2+3F\tau_{xz}{}^2,\end{aligned} \tag{31}$$

where A, B, C, D, E, F are coefficients of anisotropy which become unity in the isotropic case. As f multiplied by any constant is just as satisfactory a loading function, one of the coefficients is not needed and can be taken as unity. The form (31) is employed purely for the sake of symmetry not generality.

The stress-plastic strain relations $d\epsilon_{ij}{}^p=G(\partial f/\partial\sigma_{ij})df$ become

$$\begin{aligned}d\epsilon_x{}^p&=G(\partial f/\partial\sigma_x)df=Gdf[A(\sigma_x-\sigma_y)-C(\sigma_z-\sigma_x)]\\d\epsilon_y{}^p&=G(\partial f/\partial\sigma_y)df=Gdf[B(\sigma_y-\sigma_z)-A(\sigma_x-\sigma_y)]\\d\epsilon_z{}^p&=G(\partial f/\partial\sigma_z)df=Gdf[C(\sigma_z-\sigma_x)-B(\sigma_y-\sigma_z)]\\d\epsilon_{xy}{}^p&=1/2d\gamma_{xy}{}^p=G(\partial f/\partial\tau_{xy})df=Gdf3D\tau_{xy}\end{aligned} \tag{32}$$

etc.

The ratios of the constants may be determined by simple tests. For example, under uniaxial tension, σ_z,

$$d\epsilon_y{}^p/d\epsilon_z{}^p=A/C, \quad \text{or} \quad d\epsilon_y{}^p/d\epsilon_x{}^p=-A/(A+C). \tag{33}$$

As f is quadratic the effective stress $\bar{\sigma}$ must be taken as $c(f)^{\frac{1}{2}}$ just as for the isotropic J_2. However, the choice of c which makes $\bar{\sigma}=\sigma_x$ for uniaxial tension in the x-direction does not make $\bar{\sigma}=\sigma_y$ for uniaxial σ_y.

A definition of effective strain corresponding to the anisotropic loading function (31) is obtained quite easily by either method considered. Here, in this special form, as in the J_2 isotropic quadratic theory, $dW^p=\bar{\sigma}d\bar{\epsilon}^p$ defines a $d\bar{\epsilon}^p$ in terms of the plastic strain increments alone. This is true because $dW^p=\sigma_{ij}d\epsilon_{ij}{}^p=G2fdf=(G(2/c)(f)^{\frac{1}{2}}df)\bar{\sigma}$, and $d\bar{\epsilon}^p=G(2/c)(f)^{\frac{1}{2}}df$ is a quadratic form in $d\epsilon_{ij}{}^p$. In engineering notation Eqs. (32) when solved for the stress differences and shearing stresses become

$$\begin{aligned}\sigma_x-\sigma_y&=\frac{Bd\epsilon_x{}^p-Cd\epsilon_y{}^p}{Gdf(AB+BC+AC)};\ \text{etc.}\\\tau_{xy}&=\frac{1}{2}\frac{d\gamma_{xy}{}^p}{Gdf3D};\ \text{etc.}\end{aligned} \tag{34}$$

Substituting (34) in (31) gives f as a quadratic function of the strains divided by $(Gdf)^2$. The effective strain

[17] R. Hill, Proc. Roy. Soc. A193, 281–297 (1948).

increment is thus a constant times the square root of $f(Gdf)^2$. Noting that

$$(d\epsilon_x^p)^2+(d\epsilon_y^p)^2+(d\epsilon_z^p)^2=-2d\epsilon_x^p d\epsilon_y^p-2d\epsilon_y^p d\epsilon_z^p-2d\epsilon_z^p d\epsilon_x^p \quad (35)$$

gives

$$d\bar{\epsilon}^p=2G(f)^{\frac{1}{2}}df/c$$

$$=\frac{(2)^{\frac{1}{2}}}{c}\left\{\frac{B(d\epsilon_x^p)^2+C(d\epsilon_y^p)^2+A(d\epsilon_z^p)^2}{AB+BC+AC}+\frac{(d\gamma_{xy}^p)^2}{6D}+\frac{(d\gamma_{yz}^p)^2}{6E}+\frac{(d\gamma_{zx}^p)^2}{6F}\right\}^{\frac{1}{2}} \quad (36)$$

which reduces to the isotropic form (14) where all constants including c are unity. The constant $(2/c)$ was chosen so that

$$dW^p=\sigma_{ij}d\epsilon_{ij}^p=\sigma_{ij}G\frac{\partial f}{\partial\sigma_{ij}}df=G2fdf$$

$$=\frac{2}{c}Gc(f)^{\frac{1}{2}}(f)^{\frac{1}{2}}df=\bar{\sigma}d\bar{\epsilon}^p.$$

Equation (36) can only be used if experiments verify the hypothesis that $G=G(f)$ or the plastic work done is path independent.

For completely general quadratic anisotropic loading functions the effective strain increment is found as in Eq. (36) to be the root of a quadratic in the plastic strain increments whose coefficients are simply related to the coefficients of the stresses in f. If f is written in terms of nine stress components

$$f=A_{ij}\sigma_i\sigma_j \quad (i,j=1,2,\cdots,9), \quad (37)$$

the stress-strain relations (8) become

$$d\epsilon_i^p=2GA_{ij}\sigma_j df \quad (38)$$

which may be solved for the stresses to give

$$\sigma_i=B_{ij}d\epsilon_j^p/2Gdf \quad (39)$$

where B_{ij} is the inverse of A_{ij}; i.e., $B_{ij}A_{jk}=\delta_{ik}$. As for Eqs. (34), (36), if the plastic volume change is zero, the inverse is not unique and a variety of equivalent forms is possible. Using the same definition of effective stress and strain as before, $\bar{\sigma}=c(f)^{\frac{1}{2}}$, $d\bar{\epsilon}=(2/c)G(f)^{\frac{1}{2}}df$, gives

$$d\bar{\epsilon}^p=(1/c)(A_{ij}B_{im}d\epsilon_m^pB_{jn}d\epsilon_n^p)^{\frac{1}{2}}$$

$$=(1/c)(B_{mn}d\epsilon_m^pd\epsilon_n^p)^{\frac{1}{2}}. \quad (40)$$

If experiments show that the quadratic anisotropic f of Eq. (37) is not satisfactory for experimental correlation, loading functions of higher degree in the stresses may have to be used. As in the isotropic case, if the assumption that plastic work is path independent, $G=G(f)$, is dropped, the effective strain may well be defined by the second method in the same form (40) as for the quadratic case, simply take

$$d\epsilon_i^p\propto\frac{g(f)}{[B_{ij}(\partial f/\partial\sigma_i)(\partial f/\partial\sigma_j)]^{\frac{1}{2}}}\frac{\partial f}{\partial\sigma_i}df, \quad (41)$$

so that

$$d\bar{\epsilon}^p\propto(B_{ij}d\epsilon_i^pd\epsilon_j^p)^{\frac{1}{2}}=g(f)df. \quad (42)$$

Little more can be said about the proper forms of $\bar{\sigma}$ and $\bar{\epsilon}^p$ as experimental data is very meager. It is hoped that this type of approach will serve as a useful guide in future experimental research.

CONCLUSION

Definitions of effective stress $\bar{\sigma}$ and effective strain $\bar{\epsilon}^p$ which offer the possibility of better understanding of tests on isotropic and anisotropic work hardening materials are given on the basis of stress-strain relations from the recent "stress" theories of plasticity.

The effective stress $\bar{\sigma}$ is defined in terms of the loading function f associated with the given material. A comparison is made between two fundamentally different definitions of the effective strain $\bar{\epsilon}^p$. One, now in use, assumes that the work dissipated depends on the final state of stress alone and not on the path, and the other presented here analyzes the stress-strain relations. For $f=J_2$ the two resulting definitions of $\bar{\epsilon}^p$ coincide and, for certain other loading functions, no matter how complicated the path of loading, agree within practical limits. For many purposes, therefore, the definition of effective strain given by Dorn

$$\bar{\epsilon}^p=\int_{\epsilon_p}(d\epsilon_{ij}^pd\epsilon_{ij}^p)^{\frac{1}{2}}$$

is entirely satisfactory even though $f=f(J_2,J_3)$. Fundamental differences exist, however, and the basic work assumption must be further examined as pertinent experimental data is obtained.

Although the several $\bar{\epsilon}^p$ and $\bar{\sigma}$ as defined here will correlate a variety of tests in a useful manner, the concepts of effective strain and stress must break down for loading paths in which Bauschinger and similar effects are present.

Formation of Crystal Nuclei in Liquid Metals

D. Turnbull
Research Laboratory, General Electric Company, Schenectady, New York
(Received April 13, 1950)

The known facts about nucleation phenomena in liquid metals are interpreted satisfactorily on the basis of the critical size and interfacial energy concepts. In large continuous masses nucleation is almost always catalyzed by extraneous interfaces. However, in very small droplets the probability that a catalytic inclusion is present is so much less that their minimum nucleation frequencies are reproducible and form a consistent set of values.

Interfacial energies, σ, between crystal nuclei and the corresponding liquids have been calculated from nucleation frequencies of small droplets on the basis of the theory of homogeneous nucleation. Energies of interfaces, σ_0, one atom thick and containing N atoms were calculated from the σ's. The ratio of σ_0 to the gram atomic heat of fusion, ΔH_f, was approximately 0.45 for most metals but $\sim$0.32 for H_2O, Bi, Sb, and Ge.

The effect of relative complexity of crystal structure upon the supercooling behavior of pure metals apparently is a reflection of its effect upon ΔH_f.

THE solidification of pure metals may be thought of as a sequence of two steps: nucleation of crystals and their subsequent growth. It is known that the growth of large metal crystals into their supercooled melts is very rapid and usually limited only by the rate at which the heat of solidification can diffuse from the interface. On the other hand the time preceding the appearance of "macroscopic" crystals may be very long relative to the duration of rapid growth and often controls the solidification rate. This time delay is known as the "nucleation" period. The reciprocal of the average nucleation period is the nucleation frequency or rate and is proportional to the volume of liquid or to the area of interfaces in the system that catalyze nucleation.

Nucleation rates measured on large continuous liquid metal masses are not reproducible. However, recent experiments[1,2] on the supercooling of small metal droplets have shown that the temperatures at which the rate of crystal nucleation becomes appreciable in them are consistent and reproducible to within about ± 5 percent.

Generalizations about the solidification behavior of small droplets have been made in a preceding paper.[2] By combining these generalizations with the known facts of the supercooling behavior of large masses, a theory of nucleation can be constructed that accounts qualitatively for the nucleation behavior of liquid metal masses of all sizes and surface conditions.

CRYSTAL NUCLEATION AND ITS CATALYSIS

In nearly all theories of homogeneous (non-catalyzed) nucleation of crystals in liquids it is supposed that the nucleation period is the time necessary for a crystal of a certain critical size to form by thermal fluctuations. The radius of a critical size nucleus is directly proportional to the interfacial energy between the liquid and crystal and inversely proportional to the difference in free energy between liquid and crystal phases of infinite volume. The free energy of a crystal of critical size is decreased by fluctuations that either add or subtract atoms to it; consequently, any crystal that is smaller than the critical size will usually disappear and any that becomes larger will usually grow.

On the basis of these concepts and absolute rate theory, the following expression has been derived[3] for the homogeneous nucleation frequency of crystals in supercooled liquids:

$$I = A \exp[-K\sigma^3/(\Delta F_V)^2 kT], \qquad (1)$$

where $A = n(kT/h) \exp[-\Delta F_A/kT]$ and the other symbols are defined as follows: I = number of nuclei/sec. $\times$cm^3, n = number of atoms in the mass of liquid, K = a factor determined by the shape of the nucleus, σ = interfacial energy/cm^2 between liquid and crystal, ΔF_V = difference in free energy/cm^3 between crystal and liquid phases of infinite volume, and ΔF_A = free energy of activation for transporting an atom across the liquid-crystal interface.

In Eq. (1) three parameters σ, K, and ΔF_A are not known. However, experience indicates that ΔF_A is very small, at least for large crystals, and there is reason to believe that it is of the same magnitude as the activation energy for viscous flow, $\Delta F_A'$. If $\Delta F_A \approx \Delta F_A'$ then $\exp(-\Delta F_A/kT)$ is of the order of 10^{-2} at the solidification temperature for most metals. Therefore, it is expected that $A \approx 10^{33\pm1}$ sec.$^{-1}$ cm^{-3}. A can be evaluated from measurements of the homogeneous nucleation frequency as a function of temperature and its value compared with the predictions of theory. If the theory is thus approximately verified, Eq. (1) can be used to calculate σ (assuming some shape for the nucleus) from the nucleation frequency measured at a single temperature.

There is convincing evidence that nucleation of crystals in large continuous masses of liquid metals is almost always catalyzed by accidental inclusions and container walls.[4-6] The theory of nucleation catalysis is

[1] D. Turnbull, J. Metals 188, 1144 (1950).
[2] D. Turnbull and R. E. Cech, J. App. Phys. 21, 804 (1950).
[3] D. Turnbull and J. C. Fisher, J. Chem. Phys. 17, 71 (1949).
[4] W. T. Richards, J. Am. Chem. Soc. 54, 479 (1932).
[5] D. Turnbull, J. Chem. Phys. 18, 198 (1950).
[6] J. H. Hollomon, A.S.M. Symposium on Thermodynamics, Cleveland (1949).

also based upon the critical size concept. When crystals form on the surfaces of catalytic interfaces the total increase in free energy due to forming the crystal surfaces is less than the increase that would accompany the formation of the same mass of crystal without the aid of extraneous interfaces. Catalysis requires that interfaces be present such that the equilibrium contact angle, θ, made by the crystal and the interface in the presence of liquid be less than 180°. An expression can be written for the frequency of heterogeneous (catalyzed) nucleation analogous to the expression for homogeneous nucleation as follows:[7]

$$I_s = A' \exp[-K\sigma^3\{f(\theta)\}/(\Delta F_V)^2 kT], \qquad (2)$$

where $A' = n_s(kT/h)\exp[-\Delta F_A/kT]$, $f(\theta) = (2+\cos\theta)\times(1-\cos\theta)^2/4$, I_s = nuclei/sec. cm^2, n_s = number of atoms in the interface/cm^2, and A' is estimated to be $\approx 10^{25\pm1}$ sec.$^{-1}$ cm^{-2}. Thus a clear distinction can be made between homogeneous and heterogeneous nucleation on the basis of the coefficient A evaluated from data on the nucleation frequency as a function of temperature. However, in large masses of liquid the nucleation frequency usually follows no regular relation such as (2) because accidental catalysts of varying degrees of effectiveness (i.e. different θ values) are liable to be present in different samples. Due to these difficulties adequate testing of the nucleation theory with data on large masses of liquid metals is almost impossible.

One approach to the problem of measuring the rate of homogeneous nucleation is to pass the liquid through fine filters prior to the rate studies in order to remove catalytic particles. It is known that the frequency of nucleation in filtered samples often is much smaller than in unfiltered ones.[8,9] However, the method is not promising for the investigation of nucleation in metal samples. Because of the relatively rapid growth rate of metal crystals, a nucleus originating at a single catalytic site can quickly transform a very large mass of liquid.

A method of eliminating the effects of nucleation catalysis that appears more promising for liquid metals is to break the sample into a number of non-communicating droplets that is large in comparison with the number of catalytic sites. This technique has recently been used to study the solidification behavior of a large number of pure metals.[1,2]

Data on the isothermal rates of solidification of small droplet aggregates of tin[10] and mercury[11] at several temperatures are now available. The temperature dependence of the nucleation frequency is so great that it was easily measurable only in narrow ranges of temperature corresponding to 59 to 63° supercooling for mercury and 100 to 115° for tin. Analysis of the data showed that $A = 10^{30.4}$ for tin and $10^{30.9}$ for mercury. Considering the theoretical and experimental uncertainties these numbers are in good agreement with the value for $A \approx 10^{33}$ calculated from the theory of homogeneous nucleation but considerably larger than the value 10^{25} to be expected if nucleation were catalyzed by a film coating the entire surface of all the droplets. Thus it appears that the conditions for homogeneous nucleation were closely approached in these experiments.

Because of the sharp temperature dependence of the nucleation frequency shown in these experiments (for mercury, I changed by a factor of 10 in 1½°C), it is possible to specify a narrow range, δ, of supercooling ΔT_- such that the nucleation frequency is practically zero when $\Delta T_- < \delta$ and immeasurably fast when $\Delta T_- > \delta$. On the basis of these results it is expected that crystals would be formed at a measurable rate in small droplets of other metals within a comparatively narrow temperature band characteristic of the metal. Microscopic observations of the solidification of small droplets (10 to 100 micron diameter) of many other metals confirmed this expectation.[2]

A maximum supercooling $(\Delta T_-)_{max}$ was found for each pure metal corresponding to a temperature at which the nucleation frequency becomes appreciable. Not all droplets supercool as much as $(\Delta T_-)_{max}$ but under suitable conditions a large fraction do. For most metals $(\Delta T_-)_{max}$ is of the order of 0.18 times the absolute melting temperature, T_0. Also, it is found that $(\Delta T_-)_{max}$ is reproducible (±5 percent) and not dependent upon the source of the metal in cases where this factor was varied.

SIZE EFFECT IN NUCLEATION

According to the critical size concept, the frequency of homogeneous nucleation of crystals, I_v, in droplets of volume v is

$$I_v = Iv, \qquad (3)$$

where I is given by Eq. (1). From the results on mercury and tin I_v may be calculated for any volume of liquid metal or alternatively the amount of supercooling (ΔT_-) at which the rate of nucleation becomes appreciable in droplets of a given size can be calculated as a function of the volume. For example, mercury droplets 4 microns in diameter were found to solidify[11] with a frequency of about 10^{-5} sec.$^{-1}$ at $\Delta T_- = 60$. From the rate equation it was calculated that nuclei should form with the same frequency in a mercury droplet 1 cm^3 in volume at a supercooling $\Delta T_- \sim 49°$. Since similar relations for the rate of nucleation are expected to be valid for other substances it is inferred that large continuous masses of liquid metals (of the order of 1 cm^3 in volume) *completely free of catalytic inclusions* and *not subjected to mechanical vibration* should supercool about

[7] D. Turnbull, A.S.M. Symposium on Thermodynamics, Cleveland (1949).
[8] J. Meyer and W. Pfaff, Zeits. f. Anorg. Chemie 217, 257 (1934).
[9] V. I. Danilov and V. Neumark, J. Exp. Theor. Phys. U.S.S.R.) 10, 942 (1940).
[10] B. Vonnegut, J. Colloid Sci. 3, 563 (1948).
[11] D. Turnbull, J. Chem. Phys. 18, 768 (1950).

0.8 of the maximum supercooling observed in small droplets $(\Delta T_-)_{max}$, before solidification.

In agreement with this view, large continuous masses of liquid have been supercooled occasionally almost as much as small droplets. For example, comparatively large masses of gallium,[1] water,[12] and iron[13] have been supercooled more than 0.73 $(\Delta T_-)_{max}$.

The fact that continuous liquid masses 1 cm³ or more in volume rarely supercool more than ~ 0.05 $(\Delta T_-)_{max}$ is consistent with the viewpoint that the probability of finding accidental inclusions effective in promoting crystal nucleation is much greater in large than in small masses. This probability should increase either as the volume or surface area of the droplet so that, for example, the chances of finding one accidental inclusion in a 1 cc mass of mercury is about 10^{11} times greater than the chance of finding an inclusion in a 4 micron droplet separated from it. Thus, it appears that the known facts about the supercooling of small droplets and large masses of liquid metals are adequately accounted for on the basis of the critical size concept.

THERMAL HISTORY EFFECT IN SOLIDIFICATION

It has been fairly well established that the thermal history dependence of the nucleation frequency usually observed in the solidification of large continuous liquid masses is due to catalytic inclusions and container walls.[4,5] It follows that the nucleation frequency of small droplets should not be thermal history dependent provided that the catalysts responsible for the effect are in fact segregated in a small proportion of the droplets. This prediction has been verified by experiments on bismuth.[1] The nucleation frequency in large continuous masses of bismuth has been shown to be very dependent upon thermal history[14] but in small droplet aggregates it is independent of thermal history.[1]

From the facts that the nucleation frequency of lead crystals in sulfate-coated droplets, though increased by the catalytic effect of the film, is apparently not dependent on thermal history, it may be inferred that the presence of nucleation catalysts is a necessary but not a sufficient condition for thermal history dependence. On the basis of the micro-cavity theory of the effect,[5] the following conditions should be fulfilled in order for a thermal history effect to be observed:

1. Suitable microcavities be present in the catalytic surface. These microcavities retain small crystals above the melting temperature, but in order to be effective their diameter has to be very small (*ca.* 10^{-6} cm).
2. The substance must penetrate and fill the microcavities.
3. The contact angle θ made by the crystal with the catalytic surface when immersed in liquid should be less than 90°.

ALTERNATIVE INTERPRETATIONS OF THE SUPERCOOLING OF SMALL DROPLETS

In the foregoing it has been established that the known facts about nucleation of crystals in liquids can be explained satisfactorily on the basis of the "critical size theory." It remains to be considered whether any alternative hypotheses might explain the phenomena as well.

The basic fact to be accounted for is the long waiting period prior to the rapid growth of large crystals. Apart from the critical size hypothesis such a result might be explained if very small crystals, assumed to be stable, grow into supercooled melts at a rate many orders of magnitude less than do large crystals.

Two hypotheses that could lead to such an effect are thought worthy of serious consideration. First, it might be assumed that crystals need to have some type of imperfections in their surface in order to grow rapidly. Conceivably the probability that small crystals contain such imperfections might be much smaller than for

TABLE I. Summary of data on supercooling of small droplets.

Metal	T_0	Entropy of fusion $(\Delta H_f/T_0)$	$(\Delta T_-)_{max}$	Reference	$(\Delta T_-)_{max}/T_0$	$\times[(\Delta T_-)_{max}/T_0]^{\frac{1}{2}}[(T_0-(\Delta T_-)_{max}/\Delta H_f]^{\frac{1}{2}}$
Mercury	234.3	2.38	58	11	0.247	0.268
Gallium	303	4.42	76	1	0.250	0.218
Tin	505.7	3.41	105	10	0.208	0.216
Bismuth	544	4.60	90	1, 2	0.166	0.154
Lead	600.7	2.04	80	1	0.133	0.196
Antimony	903	5.28	135	2	0.150	0.154
Aluminum	931.7	2.74	130	2	0.140	0.183
Germanium	1231.7	4.94	227	2	0.184	0.177
Silver	1233.7	2.19	227	2	0.184	0.232
Gold	1336	2.27	230	2	0.172	0.222
Copper	1356	2.29	236	2	0.174	0.222
Manganese	1493	2.31	308	2	0.206	0.243
Nickel	1725	2.43	319	2	0.185	0.224
Cobalt	1763	2.08	330	2	0.187	0.239
Iron	1803	1.97	295	2	0.164	0.224
Palladium	1828	2.25	332	2	0.182	0.223
Platinum	2043	2.30	370	15	0.181	0.226
Water	273.2	5.28	39	25	0.143	0.149

[12] R. Smith-Johannsen, Science **108**, 652 (1948).
[13] Bardenheur and Bleckman, Stahl u. Eisen **61**, 49 (1941).
[14] W. L. Webster, Proc. Roy. Soc. **140A**, 653 (1933).

TABLE II.* Interfacial energies between various crystal nuclei and the corresponding liquid calculated from frequency of nucleation in small droplets.

Metal	Crystal structure	Interfacial energy σ ergs/cm²	σ_0 Cal./g atom	$\sigma_0/\Delta H_f$	σ_0/T_0
Mercury	Hexagonal	24.4	296	0.53	1.32
Water	Hexagonal	32.1	461	0.32	1.69
Gallium	Orthorhombic	55.9	581	0.436	1.91
Tin	Tetragonal	54.5	720	0.418	1.47
Bismuth	Rhombohedral	54.4	825	0.33	1.52
Lead	Face centered cubic	33.3	479	0.386	0.80
Antimony	Rhombohedral	101	1430	0.302	1.59
Aluminum	Face centered cubic	93	932	0.364	1.00
Germanium	Diamond	181	2120	0.348	1.71
Silver	Face centered cubic	126	1240	0.457	1.00
Gold	Face centered cubic	132	1320	0.436	0.99
Copper	Face centered cubic	177	1360	0.439	1.01
Manganese	Tetragonal	206	1660	0.480	1.11
Nickel	Face centered cubic	255	1860	0.444	1.08
Cobalt	Face centered cubic	234	1800	0.490	1.02
Iron	Body centered cubic	204	1580	0.445	0.88
Palladium	Face centered cubic	209	1850	0.450	1.01
Platinum	Face centered cubic	240	2140	0.455	1.05

* Heats of fusion and absolute melting points used in the computations summarized in Tables I and II are those recommended by K. K. Kelley ["Contributions to the Data on Theoretical Metallurgy. V. Heats of Fusion of Inorganic Substances," Bureau of Mines Bulletin No. 393 (1936)] with the exception of the heat of fusion of germanium (6100 cals/g atom) that was calculated by R. A. Oriani of this Laboratory from equilibrium data on the binary systems Ge-Au, Ge-Pb, Ge-Ag.

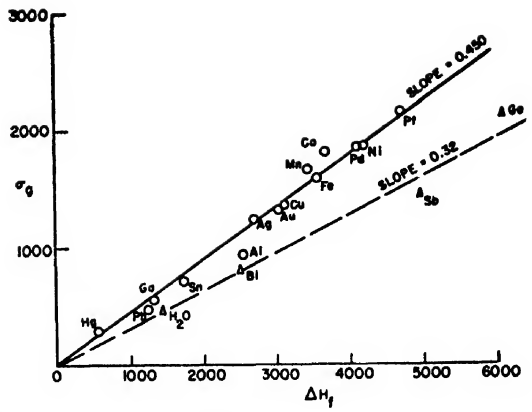


FIG. 1. Gram-atomic interfacial energy as a function of the gram-atomic heat of fusion for various substances.

large crystals so that ΔF_A would be larger for the former. However, it is not easy to construct a model that would account for such a marked effect of imperfections upon growth rate unless the interfacial energy between liquid and crystal is quite large. Thus, dissociation of the imperfection and critical size hypotheses is difficult. A further difficulty with the imperfection hypothesis is that it does not account for the effect of specimen volume upon nucleation frequency.

The other hypothesis considered is that minute amounts of soluble impurities inhibit nucleation in some way. In the experiments on the solidification of small droplets[1,2] impurity concentrations were usually of the order of 0.001 to 0.1 atomic percent. It does not seem likely that such small concentrations of impurity could inhibit nucleation by causing the liquid-crystal interfacial energy to be increased. However, it is well known that minute quantities of impurity sometimes profoundly affect the macroscopic growth rate of crystals so it is conceivable that the growth of very small metal crystals in supercooled melts is practically stopped by such impurities.

Nevertheless, there are convincing arguments against this interpretation of nucleation during solidification. A mechanism whereby impurities effectively inhibit the growth of microcrystals but not macrocrystals is possible but complicated. In the small droplet experiments $(\Delta T_-)_{max}$ was found to be independent, within experimental error, of the source and purity (within limits) of particular metals. Also it is important that the $(\Delta T_-)_{max}$ values obtained by Turnbull and Cech[2] and Mendenhall and Ingersoll[15] on the metals common to both investigations are in close agreement. This consistency in the experimental data would not be expected if the rate controlling factor were inhibition by soluble impurities. Finally it appears that the wide difference in nucleation frequency for large and small masses of liquids cannot be explained on the basis of the inhibition hypothesis unless the critical size theory is also used. There is no reason why minor amounts of soluble impurities should be more effective in inhibiting crystal growth in small droplets than in large continuous liquid masses unless it is supposed that in the latter crystals are formed by a different mechanism such that their continued growth is not inhibited. It appears that the only other reasonable mechanism involves the catalytic action of insoluble impurities.

In view of these arguments it seems that the known facts about nucleation of crystals in liquids can be interpreted much more satisfactorily on the basis of the critical size theory than in terms of either of the slow-growth hypotheses examined. The excellent correlations of the small droplet supercooling data to be presented in the following sections constitute a further decisive argument in favor of the critical size interpretation.

CORRELATIONS OF DATA ON SUPERCOOLING OF SMALL DROPLETS

The ratio of the maximum supercooling observed in small droplets to the absolute melting temperature, $(\Delta T_-)_{max}/T_0$, is nearly a constant for many metals. These ratios for 50 micron diameter droplets of various metals are given in Table I. In calculating $(\Delta T_-)_{max}/T_0$ for iron and manganese there is the problem that the metals solidify in temperature ranges in which the most stable crystal phase is different from that in equilibrium with the liquid at the normal melting point. It is not known which phase nucleates most rapidly in the liquid at the solidification temperature but it was assumed arbitrarily that the solid phase stable at the normal melting point nucleates first. The assumption can cause little error in the iron calculation since the difference in free energy between the face centered and body centered cubic phases is very small.

For all but six of the substances $(\Delta T_-)_{max}/T_0$ is

[15] C. E. Mendenhall and L. R. Ingersoll, Phil. Mag. 15, 205 (1908).

≈ 0.185 with maximum deviations of ± 0.02. Considering all the data, the ratio ranges from 0.133 for lead to 0.250 for gallium.

When the implications of the approximate constancy of $(\Delta T_-)_{max}/T_0$ are examined in terms of the theory of homogeneous nucleation (see Eq. (1)) it is found that the ratio would be constant if σ were proportional to the gram atomic heat of fusion, ΔH_f, and if the entropy of fusion $(\Delta H_f/T_0)$ were constant. Entropies of fusion of metals are approximately 2.3 cal./deg.×g atom but there are some marked deviations from this value (e.g. 4.4 for Ga). These considerations suggested that a deeper insight into the phenomena might be gained by comparing σ directly with ΔH_f.

Values of σ were calculated from the nucleation frequency at one temperature and Eq. (1) with the aid of the following assumptions:

1. Crystal nuclei are spherical in shape so that $K = 16\pi/3$.

2. $\mathrm{Exp}(-\Delta F_A/kT) = 10^{-2}$.

3. The entropy of fusion is independent of temperature.

The validity of these assumptions has been discussed elsewhere.[7] It follows from 3 that

$$\Delta F_V = \lambda \Delta T/T_0, \tag{4}$$

where λ = heat of fusion/cm^3 and $\Delta T = T - T_0$. It is believed that the available data are in general not sufficiently accurate to justify the use of a more precise free energy function. From the assumptions and Eq. (1)

$$\sigma = [\{3 \times 2.303 \lambda^2 (\Delta T)^2 kT/16\pi T_0^2\} \times \log(nkT \exp\{-\Delta F_A/kT\}/Ih)]^{\frac{1}{3}}. \tag{5}$$

I is estimated to be $10^{-(1\pm1)}$ sec.$^{-1}$ per 50-micron particle at the maximum supercooling. Although the uncertainty in I is very large, a factor of 10 error in its value introduces an error of only 1 percent in the calculated σ.

Since ΔH_f is a gram-atomic quantity it should be compared with a gram-atomic surface energy σ_g, which may be defined as the free energy of an interface containing Avogadro's number, N, atoms. If the area of such an interface is A,

$$\sigma_g = \sigma A. \tag{6}$$

Let it be assumed arbitrarily that the interface is one atom thick and let V be the gram-atomic volume; then,

$$A = N^{\frac{1}{3}} V^{\frac{2}{3}}, \tag{7}$$

where a structure factor of the order of unity is neglected. This treatment is analogous to that of Skapski[16] on liquid-vapor interfaces. Substitution of (7) into (6) gives:

$$\sigma_g = N^{\frac{1}{3}} V^{\frac{2}{3}} \sigma. \tag{8}$$

σ_g has been calculated for all of the substances whose solidification was studied by the small particle technique and the resulting values expressed in calories/gm atom are given in Table II.

Calculated ratios of σ_g to ΔH_f (see Table II) are fairly constant, ranging from about 0.31 for antimony to 0.53 for mercury. It is interesting that $\sigma_g \approx \frac{1}{2}\Delta H_f$ for one class of substances and $\approx \frac{1}{3}\Delta H_f$ for a smaller class. This difference is shown in Fig. 1 where σ_g is plotted against ΔH_f. The points for the more metallic substances (Class I) fall (within the experimental error) on a straight line of slope 0.45. Points corresponding to water and the semimetallic elements (antimony, bismuth, and germanium) (Class II) fall on a second line of slope 0.32.

Correct evaluation of σ_g depends upon the validity of the theory of homogeneous nucleation. There is an empirical correlation upon which the success of the correlation of σ_g with ΔH_f is based. This relation is:

$$[(\Delta T_-)_{max}/T_0]^{\frac{1}{3}}[\{T_0 - (\Delta T_-)_{max}\}/\Delta H_f]^{\frac{1}{3}} = M, \tag{9}$$

where M is constant for a class. M is equal to (see Table I) 0.222 ± 0.01 for Class I and 0.158 ± 0.01 for Class II.

The greater success of the $\sigma_g/\Delta H_f$ correlation relative to that of $(\Delta T_-)_{max}/T_0$ is in part illusory since the most important empirical factor (see Eq. (9)) leading to the success of the former is $[(\Delta T_-)_{max}/T_0]^{\frac{1}{3}}$. However, when this is corrected for, the $\sigma_g/\Delta H_f$ correlation is still the better of the two and in addition it appears to be more promising from a theoretical standpoint.

It is noted that the point for aluminum (see Fig. 1) on the σ_g *vs.* ΔH_f curve falls considerably below the Class I line on which it would seem to belong. Probably the most reasonable explanation of this disagreement is that surface films that may have catalyzed nucleation of aluminum crystals were not completely eliminated in the experiments.[2]

INTERPRETATION OF GRAM-ATOMIC INTERFACIAL ENERGIES

It is not surprising that mercury, tin, and the metals having cubical crystal structures fall into a distinct class with respect to solidification behavior. The four substances in Class II have the property of expanding upon solidification. In addition, their crystal structures are relatively more complex than the substances of I and excepting water over certain ranges of temperature they are all semiconductors. However, gallium exhibits all these properties yet it certainly falls in Class I. It is, of course, possible that the separation into two classes is more apparent than real. Perhaps the elements of Class II did not supercool the maximum amount for some unknown reason (in order to belong to Class I bismuth droplets should supercool 130° compared to the 90° observed) or approximations necessary in calculating σ_g may have caused misleading results. Nevertheless, the separation may be intrinsic for reasons not now apparent.

The parameter σ can be identified with the interfacial

[16] A. S. Skapski, J. Chem. Phys. 16, 386 (1948).

energy between small crystal nuclei and liquid but it is important to know whether or not it corresponds to the interfacial energy between large crystals and their melts. Any such correspondence or lack of it can only be established by measurement of σ for large crystals. However, there is some indirect evidence that σ may not be markedly dependent on crystal size. The interfacial energies between liquid nuclei and vapor calculated from the nucleation rates[17, 18] measured by Volmer and Flood[19] are in good agreement with the values measured directly on extensive liquid surfaces. Apparently decreasing the size of liquid droplets to the dimensions of nuclei has little effect upon vapor-liquid interfacial energies.

Consider now the interfacial energy between adjacent crystals of widely different orientation σ_B (i.e., grain-boundary free energy). It follows that at equilibrium $\sigma_B \leqslant 2\sigma$ at the melting point of the substance. Fisher[20] has found σ_B between adjacent copper crystals of very different orientation to be about 535 ergs/cm^2 at 800°C. σ for copper solid-liquid is calculated from nucleation data to be 177 ergs/cm^2. Considering the experimental and theoretical uncertainties the numbers are not inconsistent with the condition that $\sigma_B \leqslant 2\sigma$.

It has been assumed in making the calculations that the nuclei are spherical i.e., that σ is independent of crystal orientation. There is some experimental justification of an indirect nature supporting this assumption. It has been found[21] that the microstructures of two phase lead-copper alloys equilibrated at relatively high temperatures can be interpreted satisfactorily by assuming that the liquid lead-solid copper interfacial energy is independent of the orientation of copper crystals. A number of other microstructures in two-phase alloys have been explained on the basis of analogous assumptions.

Recently it has been shown[22] that the interfacial energy between differently oriented adjacent crystals of silicon ferrite is practically independent of the orientation difference providing it is not very small or not of such magnitude that the orientations approach a twin relationship. When the orientations are very different it appears that the grain boundary should be highly disorganized so that the transition region between the grain boundary and adjacent crystal should be somewhat similar in nature to a liquid-crystal boundary. Thus, independence of grain boundary energy upon the orientation of the abutting crystals is a point in favor of the hypothesis that liquid-crystal interfacial energies for metals are isotropic.

[17] D. Turnbull and J. H. Hollomon, Sylvania Symposium "Physics of powder metallurgy," Bayside, Long Island (August, 1949).
[18] V. K. LaMer and G. M. Pound, J. Chem. Phys. 17, 1337 (1949).
[19] M. Volmer and H. Flood, Zeits. f. physik. Chemie 170A, 273 (1934).
[20] J. C. Fisher, private communication.
[21] C. S. Smith, Metals Tech. 15, T.P. No. 2385 (1948).
[22] C. G. Dunn and F. Lionetti, J. Metals 1, T.P. No. 2517 (1949).

Although the facts that have been presented do not prove that σ or σ_0 values calculated from nucleation rates can be identified positively with the free energy of extended liquid-crystal interfaces it is believed that they afford substantial credibility for this identification.

EFFECT OF CRYSTAL STRUCTURE

It is well known that large masses of liquid metals that crystallize to close packed structures rarely can be supercooled more than 3 to 10 degrees while large masses that crystallize to more complex structures often can be supercooled much more. These facts have led to the concept[9] that the failure of large masses of the former class to supercool is an inherent property that is explained by the similarity in structure between close-packed crystals and their melts.

However, in view of the evidence that crystallization in large masses of all liquid metals is catalyzed by impurities, the facts are accounted for as well when it is assumed that effective catalysts for the nucleation of close-packed crystals occur much more frequently than those effective in the nucleation of crystals having more complicated structures.

Experiments on the supercooling of small droplets seem to have verified the latter theory. Small droplets of metals such as cobalt, silver, or nickel that crystallize to a cubical close-packed structure have been supercooled about as much relative to their melting points as have droplets of substances such as bismuth or tin that crystallize to more complex structures.

The arguments in favor of the concept that liquids crystallizing to close-packed structures should supercool much less than those crystallizing to more complicated structures hinge upon the idea that because of the structural similarity between the liquid and crystal[23] the interfacial energy between them should be small. This idea is quite reasonable, but by the same argument the heat of fusion, and therefore the free energy decrease in forming close-packed crystals should be and is small relative to that accompanying the formation of non-close-packed crystals.

Since it appears that σ_0 is directly proportional to ΔH_f for most metals the effect of crystal structure upon supercooling behavior is but a reflection of its effect upon ΔH_f within one class of substances. However, additional insight into the effect of crystal structure upon σ_0 can be gained by comparison of the σ_0/T_0 values listed in Table II. These ratios are roughly in the order of complexity of crystal structure and range from 1.91 for gallium to 0.80 for lead.

FURTHER IMPLICATIONS OF SMALL DROPLET RESULTS

The relation between σ_0 and ΔH_f that has been found is analogous to similar relations between liquid-vapor

[23] C. S. Barrett, *Structure of Metals* (McGraw-Hill Book Company, Inc., New York, 1943), pp. 224–229.

interfacial energies and heats of vaporization.[16] It suggests that analogous relations exist for liquid-crystal interfacial energies of classes of substances other than those studied in these investigations and in other types of phase transformations. For example, the interfacial energy between two solid phases in solid-state transformations of a certain class might turn out to be proportional to the heat of transformation. Unfortunately, not much effort has been made to minimize possible nucleation catalysis in solid-solid reactions so it is not known to what extent the results now available have been affected by this factor. It may be possible in some instances to minimize this factor by applying the small particle technique to solid-state reactions.

Finally, it appears that one of the most promising fields of further research in phase transformations will be that of seeking what generalizations can be made about the chemical and structure relations between nucleation catalysts and the phase that is nucleated. Already considerable progress has been made in this direction by Vonnegut[24] and Schaefer[25] on the catalysis of ice nucleation.

ACKNOWLEDGMENT

The writer has benefited from conversations with Drs. J. C. Fisher and J. H. Hollomon on some of the subjects herein discussed. He has also been helped by the encouragement of Dr. J. H. Hollomon throughout the investigations on the supercooling of small droplets.

[24] B. Vonnegut, J. App. Phys. 18, 593 (1947).
[25] V. J. Schaefer, Chem. Rev. 44, 291 (1949).

Hot-Cathode Arcs in Cesium Vapor * †

RICHARD K. STEINBERG
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received April 21, 1950)

A hot-cathode arc in cesium vapor has been studied by the use of a plane Langmuir probe to investigate the potential distribution, electron temperature and electron density in the plasma. The range of arc currents and vapor pressures over which satisfactory probe measurements may be made are determined.

In some cases it was possible to maintain the discharge with a total arc drop of less than the ionization potential of cesium (3.87 volts). Potential "humps" along the arc are shown to be absent, and ionization by the high energy electrons of a Maxwellian electron velocity distribution is found unsatisfactory as a possible mechanism for positive ion production in the arc. The mechanism of successive electron collisions can account for the observed ionization on the assumption that the effective lifetime in the excited state is increased by the process of the imprisonment of resonance radiation.

I. INTRODUCTION

THE principal purpose of this work was to obtain information which would lead to a better understanding of the mechanism by which positive ions are produced in an electric arc in cesium vapor, particularly when the total voltage across the arc is less than the ionization potential of cesium. Although various explanations for low voltage arcs have been advanced, none of them seems to enjoy universal acceptance.

Of these explanations, at least four appear to be plausible. One is that the potential gradient is not uniform along the length of the arc, but instead, a potential "hump" may exist somewhere in the arc, and there may be two points or regions with a potential difference at least as great as the ionization potential. Another explanation involves the existence of oscillations in the arc. These oscillations may be of the relaxation type at frequencies of the order of kilocycles per second, or they may be, crudely speaking, bunches of electrons or positive ions vibrating in the electric fields produced by their own displacements from their equilibrium positions. The frequencies of the latter type of oscillations may be as high as thousands of megacycles per second. With oscillations in the arc, there is the possibility that potential differences greater than the ionization potential of the gas will be present even when the average total arc drop is much less. A third explanation is based on the fact that probe measurements on the plasma of an arc usually indicate that the velocity distribution of the electrons is reasonably close to Maxwellian, at least over the range of energies for which probe measurements can be made. Without attempting to give a detailed explanation of how the high electron energies are attained, this velocity distribution function is assumed to be obeyed for all energies. Then from the probability of ionization function of the gas under consideration, it is ascertained that the number of electrons with energies greater than the ionization potential is sufficient to create ions as fast as they diffuse from the plasma to the walls. The fourth explanation does not require that a single electron deliver the entire amount

* This work has been supported in part by the Signal Corps, the Air Materiel Command and the ONR.

† This paper is an abridged version of a thesis submitted in June, 1949 at the Massachusetts Institute of Technology, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Physics. It has appeared as Technical Report No. 128, Research Laboratory of Electronics, Massachusetts Institute of Technology.

TABLE I. Cesium vapor pressure and atom density.

Temperature °C	Vapor pressure mm Hg	Atom density per cc
30	0.0000023	$(7.3)10^{10}$
100	0.00054	$(1.4)10^{13}$
150	0.0089	$(2.0)10^{14}$
200	0.079	$(1.6)10^{15}$
250	0.45	$(8.3)10^{15}$
300	1.9	$(3.2)10^{16}$

of energy necessary for ionization; rather, it suggests that it is possible for an electron to deliver only enough energy to excite an atom, with the remainder of the ionization energy coming from one or more successive collisions with other electrons. For reasons to be explained, this last mentioned mechanism is believed to be the one by which the cesium vapor arc is maintained.

Only a small amount of work on cesium vapor arcs has been reported. A series of papers by Mohler[1] is probably the most outstanding.

II. THE EXPERIMENTAL TUBES

Two tubes were constructed. The first, which will be referred to as tube No. 1, contained a plane nickel cathode and a plane tantalum anode, both 2.5 cm in diameter, mounted in a glass tube with an inside diameter of 5.3 cm. The cathode and anode were mounted on a glass and quartz assembly which could be moved along the axis of the tube by tilting the tube and shaking it slightly. The cathode-to-anode separation was 12.7 cm. Both the anode and cathode were constructed with built-in tungsten heaters capable of raising their temperatures to at least 1000°C. Low resistance, flexible, copper wire was used for making connections between the movable assembly and the lead-ins through the glass envelope. Actually, for most of the measurements, the cathode heater was turned off entirely because the arc itself caused sufficient heating to maintain the required thermionic emission.

A plane tantulum probe with a diameter of 0.25 cm was installed in tube No. 1 about 1 cm from the tube wall and with the plane of the probe parallel to the tube axis. The variation in position of the probe with respect to the arc was obtained by moving the cathode-anode assembly. A small glass tube was placed over the probe lead and the edge of the probe so that only the front surface of the probe "faced" the discharge. The same material was chosen for the anode and the probe to eliminate the uncertainty in the potential measurements which would otherwise arise from the contact difference of potential between different materials. In the case of cesium vapor, this difference would be particularly difficult to evaluate because cesium condenses on various surfaces to a thickness of the order of one atomic layer even when the temperature of the surface is hundreds of degrees centigrade higher than the temperature of the cesium supply. This layer of cesium lowers the work function of the surface by as much as 3 ev in some cases. The amount of lowering is strongly dependent on the fraction of a monatomic layer present and therefore dependent on the temperature.

Although annoying when making potential measurements, the cesium layer could be put to good use at the cathode. Thermionic currents of at least 2 amp. per cm² were attainable at cathode temperatures lower than 800°C because of the lowering of the work function of the cathode surface by the cesium. A text by Reimann[2] includes a more extensive discussion of the phenomenon together with references to original literature.

The cesium supply was contained in an appendix to the tube which consisted of a glass tube about 6 in. long. The cesium temperature, and therefore the vapor pressure, was controlled by the temperature of an oven which enclosed the appendix. To prevent bulk condensation of the cesium on the glass walls where it would cause low resistance leakage paths between the various

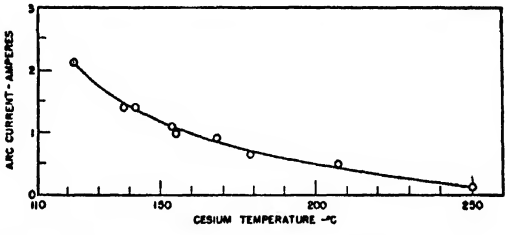


FIG. 1. Minimum arc current for non-oscillating condition as a function of cesium temperature (tube No. 1).

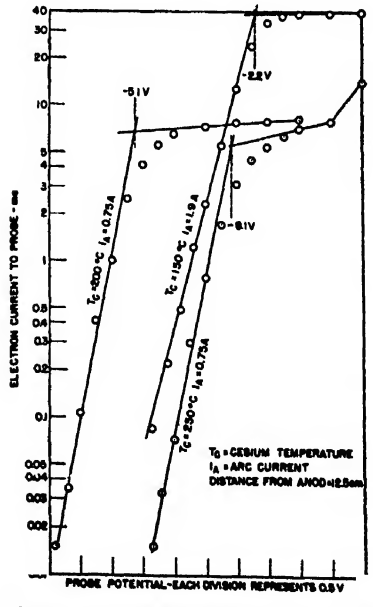


FIG. 2. Sample potential determinations (tube No. 2).

[1] F. L. Mohler, J. Research Nat. Bur. Stand. 9, 25, 493 (1932); 10, 771 (1933); 16, 227 (1936); 17, 849 (1936); 21, 697 (1938). C. Boeckner and F. L. Mohler, J. Research Nat. Bur. Stand. 10, 357 (1933).

[2] A. L. Reimann, *Thermionic Emission* (John Wiley and Sons, Inc., New York, 1934), p. 135 ff.

leads, a separate oven which could be maintained at a temperature higher than the temperature of the cesium was constructed for the main body of the tube. A glass window was installed in one side of the oven to provide a means for viewing the arc.

Tube No. 2 contained a series of probes in line as a means for measuring various plasma properties as functions of distance along the arc. In tube No. 2, the diameter of the front face of the cathode was 4.3 cm and the inside diameter of the glass tube was 4.6 cm. The anode was simply a disk 4.3 cm in diameter. The material for both the cathode and anode was tantalum. Four plane probes, each similar to the one used in tube No. 1, were mounted in a line along the length of the tube, and the distances from the center of the probes to the anode were 0.5, 2.5, 12.5, and 14.7 cm. The cathode-to-anode separation was 15.2 cm.

Both tubes were baked at 500°C for one hour, and the cathode and anode of each were heated to a bright red for several minutes either by means of their heaters or by radiofrequency induction heating. The pressure at the time the tubes were sealed off the vacuum system, as indicated by an ionization gauge, was about $(3)10^{-6}$ mm Hg. This reading corresponds roughly to the vapor pressure of cesium at room temperature and is the highest vacuum that could be expected.

III. CESIUM VAPOR PRESSURE AND ATOM DENSITY

The most recent and probably the most accurate vapor pressure data which were found were those of Taylor and Langmuir.[3] They measured the vapor pressure up to a temperature of 73°C by utilizing a method which makes use of the fact that practically every cesium atom which strikes a hot tungsten surface leaves as a positive ion. From their data, they derived the following formula for the vapor pressure of cesium in equilibrium with the liquid state:

$$\log_{10} p = 11.0531 - 1.35 \log_{10} T - (4041/T), \qquad (1)$$

where p is the pressure in millimeters of mercury, and T is the temperature of the cesium in °K. They claim an accuracy of 3 percent up to 327°C and 8 percent up to 727°C when extrapolating the curve beyond the range over which data were taken. Table I shows the cesium vapor pressure and atom density at various temperatures as determined from Eq. (1) and the equation of state.

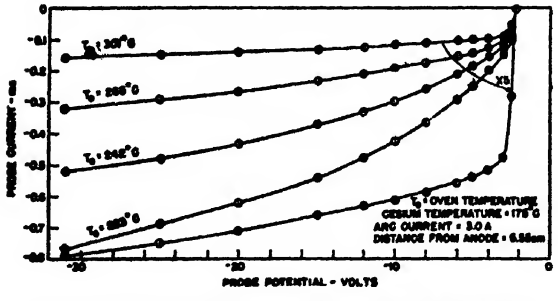


FIG. 3. Negative current portion of probe characteristic curves (tube No. 1).

IV. PLASMA POTENTIAL

The usual Langmuir technique was used for making measurements of the plasma potential. A text by Loeb[4] includes a discussion of the method and its limitations.

It was found that there was a rather limited range of arc currents over which reliable probe measurements could be made. Oscillations with frequencies in the range from 1000 to 10,000 c.p.s. would occur in the arc when the arc current was reduced below a certain critical minimum value. This minimum was a function of vapor pressure (cesium temperature) and was lower at the higher pressures, as shown in Fig. 1. The amplitude of these oscillations depended upon several parameters including vapor pressure, cathode temperature and arc current, but they usually had an amplitude of the order of 1 volt as indicated by an oscilloscope connected across the anode and cathode. The oscillations could not be eliminated by circuit modifications. Probe data were taken only at those values of arc current for which no oscillations existed.

The upper limit to the arc current at which reliable probe measurements could be made was set by the fact that at the higher arc currents the probe characteristic curve did not have a well-defined bend in it, or it had no bend at all, at plasma potential. Instead, the probe current would continue to rise as the probe voltage was increased until the probe was drawing the entire arc current. Also, the probe currents obtained at the higher arc currents were sufficient to cause probe heating, which, because of the cesium layer, changed the work function of the probe and introduced another source of error in the measurements. With tube No. 1, good probe characteristic curves were obtained with arc currents up to 3 amp. and up to 1.9 amp. with tube No. 2. The limitations just described were such that no reliable probe measurements at all could be made at cesium temperatures below about 150°C. Measurements were not made at cesium temperatures higher than 250°C because of possible damage to the tubes by chemical attack by cesium.

Sample logarithmic plots of probe characteristic curves are shown in Fig. 2.

Probe data taken at potentials negative with respect to floating potential were found to be of little value. Figure 3 shows the negative portion of the probe characteristic curves taken with the plane probe in tube No. 1 when the cesium temperature and arc current were held constant at 175°C and 3 amp. respectively, but when the main oven was at various temperatures from 223°C to 301°C. The lower oven temperatures appear to cause a much greater positive ion current to the probe. Similar sets of curves were obtained at cesium

[3] J. B. Taylor, I. Langmuir, Phys. Rev. 51, 753 (1937).

[4] L. B. Loeb, *Fundamental Processes of Electrical Discharge in Gases*, (John Wiley and Sons, Inc., New York, 1939), pp. 232–257.

temperatures of 125°C and 250°C. It was discovered that even when the oven temperature was much higher than the cesium temperature, enough cesium would condense on the glass surfaces in the tube to cause leakage which, although small, was great enough to cause substantial errors in the measurement of positive ion currents. Thus, since the plane probe was doubtlessly in contact with its glass shield, the shield could have acted as a part of the probe, with the effect of increasing its area. The increase in leakage resistance with temperature would then account for the smaller probe currents found at the higher oven temperatures. The negative probe currents that were obtained at an oven temperature of 300°C were possibly a reasonably good measure of the positive ion current present, but, again because of the cesium layer on the surface, thermionic emission from the probe surface is of the same order of magnitude at this temperature.

From kinetic theory the relationship

$$I = n(kT/2\pi M)^{\frac{1}{2}} \qquad (2)$$

may be obtained, where I is the random current of gas particles crossing unit area in unit time, n is the number of particles per unit volume, k is Boltzmann's constant, T is the absolute temperature and M is the mass of the particle. This equation indicates that if the electron and positive ion densities and temperatures were equal, the ratio of random electron current to random positive ion current would be equal to the square root of the ratio of the ion to electron masses. The square root of this ratio is approximately 500 in the case of cesium. In order to determine plasma potential more accurately, the positive ion current was assumed to be 1/500 of the random electron current and this correction was applied to the measured values of probe current. The logarithmic plots so obtained (of which those in Fig. 2 are examples) were substantially linear in the region of rapidly varying electron current to the probe. The utilization of the data obtained at probe potentials more negative than floating potential was not attempted except to confirm, as near as could be ascertained, the fact that 500 was a reasonable value for the electron-ion current ratio.

Experimental work on mercury vapor by Robert Howe at the Massachusetts Institute of Technology, and by others, has indicated that the random electron-ion current ratio is actually less than the square root of the mass ratio in spite of the fact that the electron temperature is usually found to be quite high. In the plasma potential determination, a considerable variation in the ratio which is used causes only a small variation in the potential measurement.

TABLE II. Total arc drop (tube No. 1)

Cesium temperature °C	Arc current amp.	Arc drop volts
150	3.0	3.0– 9.0
200	3.0	6.6– 7.8
200	1.0	6.0–10.4
250	3.0	5.5–10.7
250	1.0	8.2–13.5
250	0.3	21.5

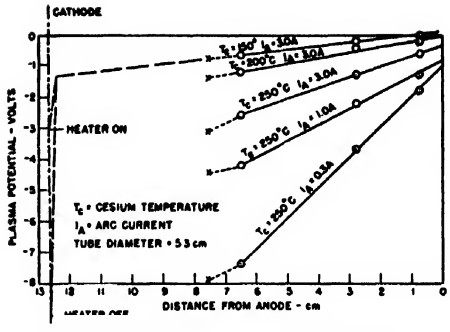


FIG. 4. Plasma potential (tube No. 1).

For tube No. 1 the plasma potential, referred to the anode as a function of distance from the anode, is shown plotted in Fig. 4 for various cesium temperatures and arc currents. It was intended to take data with the probe at positions close to the cathode but the movable assembly became damaged before this was accomplished. However it happened to be possible to take some data using the tantalum electrode as the cathode and the nickel electrode as the anode, with a cathode-to-anode separation of 8.5 cm and with the probe 0.9 cm from the cathode. The data so taken are represented by crosses and dotted lines in Fig. 4. The indication is that the plasma potential is at least a reasonably linear function of distance to within 0.9 cm from the cathode. Most of the data were taken with the cathode heater current turned off entirely and the arc itself was used as a source of cathode heating power. In order to obtain electrons for starting the arc it was necessary to use the cathode heater.

The total arc drop was a function of cathode temperature, and by using the heater, this total arc drop could be reduced to a value several volts below the value obtained without the heater. Excessive heating caused the cesium layer to evaporate faster than it was deposited, so that in spite of the increased temperature, the resulting increase in work function of the cathode surface caused a reduction in the available thermionic emission and an increase in arc drop. At a cesium temperature of 150°C and an arc current of 3 amp., for example, the total arc drop was 9 volts with the heater off. By using the heater, the drop could be reduced to a value as low as 3 volts. This example is shown graphically in Fig. 4.

The results, to be described later, obtained with tube No. 2 indicate that a linear extrapolation of the plasma potential to within 0.5 cm of the cathode is justified. Table II indicates the arc drops which were recorded at the times the probe data were taken.

With the tantalum cathode and nickel anode, the arc drop could be made as low as 2 volts. It was possible, though difficult, to reproduce these minimum values; the difficulty arose from the fact that it was necessary

that a large portion of the cathode be at the optimum temperature for thermionic emission and that the cathode heating by the arc tended to make the temperature non-uniform. The important point here, however, is that the potential of any point in the plasma when referred to the anode was found to be independent of cathode temperature and total arc drop. Several readings of potential were made under different conditions of arc drop and in each instance, even though the arc drop varied as much as 6 volts (the largest variation recorded), the potential of the plasma with respect to the anode stayed the same within experimental error, which was estimated to be about 0.1 volt. If the 150°C-3 amp. curve is considered again, it is seen that the negative anode fall is no greater than about 0.2 volt, and that with a total arc drop of only 3 volts there are no

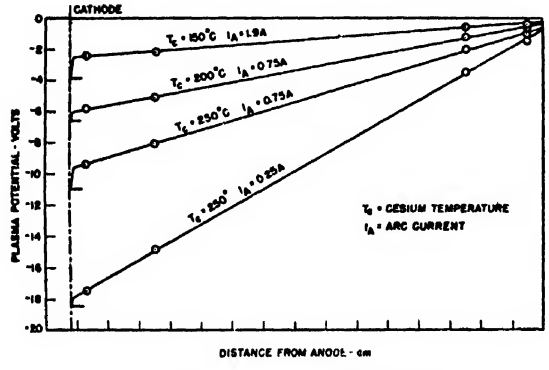


FIG. 5. Plasma potential (tube No. 2).

two points in the discharge with a potential difference as great as the ionization potential of cesium (3.87 volts).

The plasma potential measurements obtained from tube No. 2 are shown plotted in Fig. 5. It can be seen that they agree both qualitatively and quantitatively with those obtained from tube No. 1 if consideration is given to the fact that the cross-sectional area of tube No. 2 was only 0.75 times as great as that of No. 1.

Most of the data taken with tube No. 2 were taken with the cathode heater turned on and adjusted to yield minimum obtainable total arc drop. The total arc drop is indicated for each curve at the left edge of Fig. 5. The curve corresponding to a cesium temperature of 150°C and an arc current of 1.9 amp. shows no indication that there are two points in the tube with a potential difference greater than the ionization potential of cesium, a result similar to that obtained from tube No. 1. Cathode falls of from 0.8 to 1.5 volts were found.

The total arc drop as a function of arc current for tube No. 2 at the three vapor pressures studied is shown plotted in Fig. 6. For each curve the points represent the lowest drop which was attainable with the particular cathode which happened to be in use. For the higher arc currents, i.e., above about 3 amp., the heating of the cathode by the arc itself was sufficient to cause the cathode temperature to rise above optimum value for thermionic emission. Therefore it was necessary, in obtaining these data, to measure the arc drop immediately after starting the arc. Within a few seconds the arc drop would rise by as much as a volt or more above its initial value. If the cathode had been capable of emitting even greater thermionic currents the total arc drop might have been less because there was no apparent lower limit to the cathode fall other than zero.

The random electron current as a function of distance along the arc for two typical cases is shown in Fig. 7. The random current near the cathode was high, as shown, when the cathode heater was used to enable the cathode to emit a copious supply of thermionic electrons. When the heater was turned off, making it necessary for the arc itself to heat the cathode through the high cathode fall thus created, the random electron current near the cathode was as much as an order of magnitude less. This effect was in spite of the fact that no measurable difference in the space potential was noted.

An indication of the electron temperature may be obtained from the slope of the curve representing the logarithm of the electron current to the probe as a function of probe potential. Actually, the term "temperature" does not have a well-defined meaning when applied to the particles in the plasma of an arc; nevertheless, the temperature obtained in this manner is plotted as a function of distance along the arc in Fig. 8. Both of these sets of data were obtained with the four plane probes in tube No. 2. Similar results were obtained with tube No. 1.

V. OSCILLATIONS

Except for the low frequency oscillations, previously mentioned, which occurred when the arc current was reduced below a certain critical value, no oscillations were found. During all of the runs an oscilloscope which was responsive to frequencies up to 2 Mc was connected across the anode and cathode of the tube to check for the existence of oscillations, and data were taken only when no oscillations were indicated.

A search was made under various conditions of arc current and vapor pressure for oscillations at frequencies higher than might have been indicated on the oscilloscope. To carry out the search, a radio communications receiver, a radar search receiver and a crystal detector were employed in conjunction with various tuned circuits and Lecher wires connected in the arc circuit. A failure to find oscillations, no matter how thorough the search, does not establish the fact that no oscillations were present; however it is believed to have been reasonably well established that no oscillations were present at frequencies below an upper limit of about 10,000 Mc/sec. According to a derivation given by Tonks and Langmuir,[5] it is possible for electron oscillations in a plasma to occur at a frequency of $(8980)n^{\frac{1}{2}}$ where n is the number of electrons/cm³. The highest electron density found in the range of arc currents and

[5] L. Tonks, I. Langmuir, Phys. Rev. 33, 195, 990 (1929).

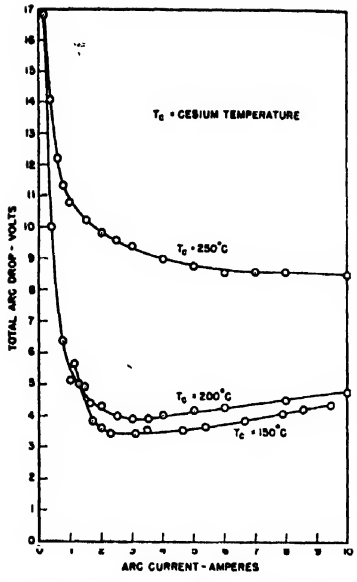


FIG. 6. Arc voltage *vs.* arc current (tube No. 2).

vapor pressures over which data were taken was $(1.6)10^{12}$ electrons/cm³. This density was obtained near the cathode of tube No. 2 when the cesium temperature was 150°C and the arc current was 1.9 amp., and it would imply a frequency of 12,000 Mc which might have been higher than the crystal was capable of detecting. However, 12,000 Mc is an upper limit to the possible frequency, and even under the same arc conditions the electron density near the anode was $(1.6)10^{11}$ electrons/cm³. This density would indicate a frequency of 3600 Mc, which was known to be within the range over which the detecting instruments would function.

High frequency oscillations have been reported occasionally in connection with tubes containing arcs, and certainly can be obtained with tubes of special design. The existence of the oscillations so found is usually not assured by the mere presence of the arc but is critically dependent on such factors as arc current and potentials of the various electrodes that may be present in the tube.

It is believed that there is little evidence to support a theory of ionization based on oscillations, particularly since, even if their existence were established with certainty, it would yet remain to be shown that they had sufficient amplitude to produce ionization. On the other hand, the present failure to detect oscillations is an indication that the ionization could not have been dependent upon them.

VI. IONIZATION BY ELECTRONS WITH A MAXWELLIAN VELOCITY DISTRIBUTION

The number of ions lost to the walls per centimeter-length of tube may be obtained experimentally and is given by the expression $2\pi Ri_+/e$, where R is the radius of the tube in centimeters, i_+ is the positive ion current in amp./cm² as determined by probe measurement, and e is the electronic charge expressed in coulombs. For the range of atom density for which the discharge fills the whole tube, the number of ionizing collisions per electron per sec. is then

$$\alpha = 2\pi Ri_+/e\pi R^2 n = 2i_+/eRn, \qquad (3)$$

where n is the number of electrons per cm³, and, for the examples to be considered, it will be assumed that n is not strongly dependent on distance from the axis of the arc. If their velocity distribution is Maxwellian, the number of electrons per cm³ with velocities between v and $v+dv$ cm per sec is

$$nf(v)dv = 4\pi n\left(\frac{m}{2\pi kT_e}\right)^{\frac{3}{2}} v^2 \exp\left(-\frac{mv^2}{2kT_e}\right)dv, \qquad (4)$$

where m is the mass of the electron in grams, k is Boltzmann's constant expressed in ergs per degree, and T_e is the electron temperature in °K. The number of ionzing collisions per sec. that will be produced by an electron of velocity v is (PvN/N_0) where N is the number of atoms per cm³ present, N_0 is equal to $(3.5)10^{16}$ and is the number of atoms per cm³ when the temperature and pressure are 0°C and 1 mm Hg respectively, and P is the probability of ionization per centimeter-length of path when N_0 atoms per cm³ are present. Note that P is the product of the atomic cross section for ionization and N_0. The number of positive ions per cm³ produced per second by electrons with velocities between v and $v+dv$

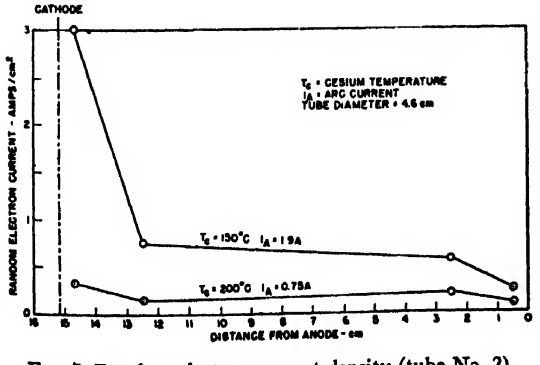


FIG. 7. Random electron current density (tube No. 2).

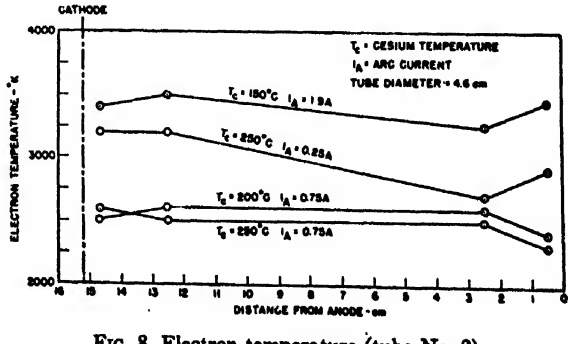


FIG. 8. Electron temperature (tube No. 2).

is then

$$4\pi n\left(\frac{m}{2\pi kT_e}\right)^{\frac{1}{2}} P\frac{N}{N_0} v^3 \exp\left(-\frac{mv^2}{2kT_e}\right)dv. \quad (5)$$

If the substitution

$$v=(2Ve10^7/m)^{\frac{1}{2}} \quad (6)$$

where V is the electron energy expressed in electron volts, is made, expression (5) takes the form

$$(8)10^{14}\pi n\left(\frac{m}{2\pi kT_e}\right)^{\frac{1}{2}} \frac{e^2}{m^2} P\frac{N}{N_0} V \exp\left(-\frac{eV10^7}{kT_e}\right)dv. \quad (7)$$

The probability of ionization for cesium is not known, but it will be assumed for the purpose of this calculation that it is zero for electron energies less than the ionization potential V_i, and is constant with a value P_i, at all energies above V_i. After dividing by n, an integration of expression (7) from $V=V_i$ to $V=\infty$ yields the number of positive ions α' formed per second per electron. The result is

$$\alpha'=(2.0)10^{-7}\frac{P_iN}{T_e^{\frac{1}{2}}}\exp\left[-\frac{(1.16)10^4V_i}{T_e}\right] \times\left(V_i+\frac{T_e}{(1.16)10^4}\right), \quad (8)$$

where V_i is in volts.

By using the data obtained from tube No. 2 when the cesium temperature was 150°C and the arc current was 1.9 amp., for example, it was found that α, as computed from Eq. (3), was approximately 16,700 ionizing collisions per electron per sec. Although i_+ and n vary along the length of the plasma, they vary in approximately the same fashion so that α is reasonably constant. As described previously, i_+ was taken as 1/500 of the random electron current. Upon using 3.87 volts for V_i and 3400°K for T_e, it was found by equating α to α' that P_i must be at least 3000 ions per electron per cm of path. Data for the probability of collisions taken by Brode[6] indicate that it is 650 collisions per cm at V_i and drops off considerably at higher electron energies. Brode's probability of collision includes elastic and exciting collisions as well as ionizing collisions. From these results it is concluded that the number of electrons with sufficient energy for ionization which is made available from the Maxwellian distribution is not adequate to sustain the plasma. The results of additional examples are shown in Table III.

By working along similar lines Killian[7] concluded that, for mercury vapor, the ionization produced by this mechanism was sufficiently great in some cases to make up for the loss of ions caused by diffusion to the walls. However he obtained values for electron temperature up to 38,000°K, which is rather high. Because of the exponential term, α' is a rapidly varying function of temperature, and if the true temperature were even slightly less the mechanism would have been found insufficient in all cases.

[6] R. B. Brode, Phys. Rev. 34, 673 (1929).
[7] T. J. Killian, Phys. Rev. 35, 1238 (1930).

TABLE III. Data taken from tube No. 2.

Cesium temperature, °C	150	200	250	250
Atom density, per cc	$(2.0)10^{14}$	$(1.6)10^{15}$	$(8.3)10^{15}$	$(8.3)10^{15}$
Arc current, amp.	1.9	0.75	0.75	0.25
Electron temperature, °K	3400	2600	2500	3000
Electron density, per cc	$(4.8)10^{11}$	$(1.4)10^{11}$	$(9.5)10^{10}$	$(1.9)10^{10}$
Necessary P_i, per cm (see Section VI)	3000	18,000	66,000	380
Necessary P_e, per cm (see Section VII)	130	410	290	72

It should be noted that oscillations do not necessarily afford an explanation for the presence of high energy electrons or the existence of a high electron temperature, as is sometimes implied.[7,8] It is easy to see, though, why oscillations might cause probes to indicate, erroneously, high energy electrons or high electron temperatures. The explanation lies in the fact that, over a portion of the oscillation cycle, the probe at any potential setting is more positive with respect to the plasma than it would be in the absence of oscillations, and an increased electron current to the probe is produced. Because of the exponential current-probe potential relationship, the decrease in electron current that occurs over the opposite portion of the cycle does not cancel the effect. Thus at all probe potentials negative with respect to the average plasma potential, the electron current to the probe could be misleadingly large in the presence of oscillations.

An additional difficulty with the theory which assumes a Maxwellian velocity distribution of electrons lies in the inability to explain how the supply of high energy electrons is maintained in view of the fact that the number of high energy electrons would be expected to be diminished because of the inelastic (exciting and ionizing) collisions which occur. It is somewhat surprising that probe measurements indicate conformance to a Maxwellian distribution over as wide a range as they do, but this conformance is not a good reason to extrapolate the results beyond the range of measurements, which seldom extend to energies much higher than that corresponding to ionization potential. Further, measurements on the high energy electrons are subject to considerable error because of the uncertainties involved in extrapolating that portion of the probe characteristic curve which supposedly represents positive ion current.

VII. IONIZATION BY SUCCESSIVE COLLISIONS

The successive-collision hypothesis was suggested as far back as 1925,[9–11] but unless certain other phe-

[8] N. L. Oleson, C. G. Found, J. App. Phys. 20, 416 (1949).
[9] K. T. Compton and C. Eckart, Phys. Rev. 25, 139 (1925).
[10] W. Fabrikant, Techn. Phys. U.S.S.R. 5, 864 (1938).
[11] B. Klarfeld, Techn. Phys. U.S.S.R. 5, 913 (1938); Comptes Rendus de l'Acad. des Sciences U.S.S.R. 24, 251 (1939).

nomena are taken into consideration, which was not always done, ionization by successive collision appears unlikely. The reason is that the lifetime of an excited state is so short compared with the average time between collisions that there is only a small probability that an excited atom will still be excited by the time a second electron strikes it. If the case of an arc current of 1.9 amp. and a cesium temperature of 150°C is taken as an example, and if Eq. (8), which assumes zero probability at energies below a certain critical value and a constant probability at higher energies, is employed, it is found that the number of excited atoms created per second per electron is $(8.1)10^3P_e$. The probability of excitation is P_e and the first critical potential of 1.48 volts was used. If $(4.8)10^{11}$ electrons per cm^3 is taken as the electron density and if the average lifetime of an excited state is 10^{-8} sec., it is found that the density of excited atoms is approximately $(3.9)10^7P_e$ per cm^3, which is about $(1.9)10^{-7}P_e$ times the density of unexcited atoms. Since an additional 2.39 ev are required to ionize an atom which has already been excited to the lowest excitation level, a second application of Eq. (8) indicates that the number of ions produced per second per electron is about $(1.1)10^{-4}P_eP_i'$, where P_i' is the probability of ionizing an excited atom if all the atoms were excited. In correspondence with the units used above, the atomic cross section for ionization of an excited atom is equal to P_i'/N_0. When the result just obtained is compared with α, the necessary number of ions per sec. per electron (16,700 for this example), it is seen that the product P_eP_i' must be at least $(1.6)10^8$ which is rather high. The assumption of a Maxwellian distribution of electron energies is possibly more justified in this case because the energies under consideration are not so great.

The phenomenon of imprisonment of resonance radiation is yet to be considered, and according to Holstein[12] the lifetime of an imprisoned light quanta in cesium vapor over a pressure range of 0.01–1 mm Hg is $(1.2)10^{-5}R^{\frac{1}{2}}$ sec., where R is the radius in centimeters of an infinitely long tube. Since it makes no difference which atom is excited, the passing of resonance radiation back and forth among the atoms has the effect of increasing the lifetime of an excited state by over 10^3 in the present case.

Also, according to Klarfeld,[11] the probability of ionization for excited atoms may be at least 10 times the probability for ionization from the ground state. The higher probability seems feasible when it is remembered that the wave function of an excited atom is much more spread-out than the wave function of an atom in the ground state. Thus when these two factors are considered it is found that P_e need not be greater than the square root of $(1.6)10^4$ or approximately 130 collisions per cm. This result causes ionization by successive collisions to appear more plausible. Additional examples are included in Table III.

VIII. CONCLUSION

Various properties of hot-cathode arcs in cesium vapor have been measured. In an effort to determine the mechanism by which the plasma of the arc is maintained it has been observed that no potential "humps" exist in the plasma, and, under certain conditions, no oscillations could be found. Sufficient ionization by the high energy electrons of a Maxwellian distribution appears unlikely. If the phenomenon of imprisonment of resonance radiation is taken into consideration, it has been shown that ionization by successive collisions is a plausible mechanism for plasma maintenance in spite of the short lifetimes of the excited states.

IX. ACKNOWLEDGMENT

I wish to thank Professor W. B. Nottingham, under whose direction the work was carried out, for the many helpful suggestions received. I also wish to thank Dr. A. W. Hull of the General Electric Company Research Laboratory, Schenectady, New York, for suggesting the problem and for assistance received in the preparatory phase of the work.

[12] Alpert, McCoubrey, and T. Holstein, Phys. Rev. 76, 1257, 1259 (1949). (The information on cesium vapor was contained in a private communication from Holstein.)

Allotropy of Beryllium*

S. S. Sidhu and C. O. Henry
University of Pittsburgh, Pittsburgh, Pennsylvania
(Received April 24, 1950)

The allotropy of beryllium has been confirmed by employing different alloying properties of allotropic forms of metallic elements. Alloys of beryllium and gold, studied by x-ray diffraction methods, showed that the limit of solubility of gold in beryllium is between two and three atomic percent, and practically all of it is dissolved in the β-form. The lattice parameters of the two forms observed here are: $a_0=2.286$A, $C_0=3.588$A, and $C_0/a_0=1.570$ for the α-form, and $a_0=6.93$A, $C_0=11.35$A, and $C_0/a_0=1.638$ for the β-form. The volume of the unit cell of the β-form is about 29 times that of the α-form.

SOME investigators[1–3] have reported allotropy of beryllium while others[4] have failed to observe it. As a result a considerable uncertainty[5] exists regarding it. The object of this investigation was to obtain definite information in favor of or against such an existence.

EXPERIMENTAL PROCEDURE

The allotropic forms of metallic elements differ in their crystalline structures, and hence possess different alloying properties. These properties may be employed to prove or disapprove the allotropy of an element. If heavy atoms, such as of gold, can be dissolved, even to a limited solubility, in one of the allotropes of a light element as beryllium, noticeable changes may occur in the intensities of its diffraction lines due to a large difference in the atomic scattering factors of the two atoms. With this in view four alloys having compositions as given in Fig. 1 were prepared in an evacuated induction furnace with maximum pressure of the order of 10^{-4} mm Hg. Weighted mixtures of gold and beryllium metals were melted in BeO crucibles covered with BeO lids and enclosed in graphite heaters and ZrO_2 insulators. The maximum temperature varied between 1400° and 1480°C, and the solidification of the melts took place in the furnace. Powder x-ray diffraction patterns were made of stress relieved beryllium and representative samples of each alloy with a cylindrical camera of 114.6 mm diameter and copper $K\alpha$-radiation. The diffraction pattern of Be was repeated with a flat film and a specimen to film distance of 60 mm.

Fig. 1. X-ray diffraction photographs of: (A) α and β-forms of Be; (B) Be with 0.02 atomic percent Au; (C) Be with 0.25 atomic percent Au; (D) Be with 2 atomic percent Au; (E) Be with 3 atomic percent Au. Camera diameter, 114.6 mm; Copper K_α-radiation.

RESULTS

As given in Table I the diffraction data obtained from the pattern of a sample of beryllium contained diffraction lines of the usual form of beryllium (α-form), and in addition several weak lines at relatively small Bragg angles. The interplanar spacings corresponding to the latter lines alone were not sufficient to determine a crystal structure but they could be fitted into an hexagonal structure with Miller indices as given under $\beta-$Be, and the lattice parameters as given below. These

Table I. Estimated relative intensities (I/I_0), interplanar spacings (d), and the Miller indices (hkl) for α- and β-forms of beryllium.

I/I_0*	d(A)	hkl ($\alpha-$Be)	hkl ($\beta-$Be)
W	3.47	—	110
$V.W$	3.12	—	103
W	2.86	—	004
W	2.56	—	104
W	2.34	—	203
W	2.22	—	114
$V.W$	2.13	—	105
W	2.06	—	204
S	1.977	100	—
$V.W$	1.910	—	006
M	1.794	002	—
$V.S$	1.736	101	—
$V.W$	1.351	—	
M^+	1.328	102	—
$V.W$	1.272	—	
$V.W$	1.236	—	
S	1.143	110	—
S	1.023	103	—
W^+	0.991	200	—
S^-	0.965	112	—
M^+	0.955	201	—
$V.W$	0.911	—	
W	0.897	004	—
W	0.867	202	—
M	0.817	104	—

* V = Very; S = Strong; M = Medium; W = Weak.

* This work has been supported in part by ONR.
[1] F. M. Jaegar and J. E. Zanastra, Proc. Amst. Akad. Sci. 36, 636 (1933).
[2] G. F. Kosolapov and A. K. Trapoznikov, J. Exper. Theor. Phys. U.S.S.R. 6, 1136 (1936).
[3] G. P. Chatterjee and S. S. Sidhu, Phys. Rev. 76, 175 (1949).
[4] P. Gordon, J. App. Phys. 20, 908 (1949).
[5] L. Gold, AEC Declassified Report No. 2643 (July, 1949).

TABLE II. Estimated relative intensities and interplanar spacings for sample of Be with 2 atomic percent of gold.

I/I_0	d(A)
V. S	3.47
M−	3.03
M	2.81
W	2.45
V. W	2.33
S (double)	2.12
V. W	2.06
S	1.977*
V. W	1.910
S	1.794*
V. S	1.736*
V. W	1.639
W+	1.579
W	1.503
W+ (double)	1.376
M	1.328*
W	1.296
W	1.258
M (double)	1.222
V. W	1.173
S	1.144*
V. W	1.086
W	1.037
M	1.023*
V. W	0.990*
M	0.964*
M	0.954*
W	0.898*
W	0.865*
W	0.816*

* Lines of $\alpha-$Be.

TABLE III. Estimated relative intensities and interplanar spacings of sample of Be with 3 atomic percent gold.

I/I_0	$d_{(A)}$	$\frac{hkl}{AuBe_5}$	$\frac{hkl}{\alpha-Be}$
V. S	3.50	111	
S	3.03	200	
S	2.15	220	
W	1.977		100
S	1.831	311	
V. W	1.793		002
M	1.752	222	
M	1.733		101
W	1.520	400	
M	1.397	331	
M	1.362	420	
W	1.330		102
M	1.240	422	
M	1.173	333, (511)	
W	1.144		110
W	1.076	440	
M	1.030	531	
M	1.016	600	
V. W	0.989		200
V. W	0.965		112
M	0.958	620	
V. W	0.956		201
W	0.928	533	
W	0.919	622	
V. W	0.899		004
W	0.879	444	
V. W	0.869		202
M	0.853	551, (711)	
M	0.845	640	
M	0.815	642	104
M	0.794	533, (731)	

lines could not be attributed to impurities as determined by the spectrographic and x-ray diffraction analyses.

X-ray diffraction data obtained from patterns of alloys B, C, and D showed that there was no observable effect on the diffraction lines of the $\alpha-$Be with the increasing atomic percent of gold. The relative intensities and the interplanar spacings of these lines remained identically the same as those observed in the sample of beryllium. The diffraction lines of the $\beta-$Be on the other hand showed marked changes. The intensities of some of the original lines increased without any apparent change in their interplanar spacings, and additional lines appeared in increasing numbers as the atomic percent of gold increased. A few of the original lines appeared to be resolved into two lines. As given in Table II the diffraction data of sample D contained many more lines than those obtained from the sample of beryllium. These lines, however, were not due to gold or to the phase $AuBe_5$.

As shown in pattern *E*, Fig. 1, a definite transformation takes place at about 3 atomic percent of gold. The diffraction data, given in Table III, show conclusively that the alloy contained mostly $AuBe_5$ phase, which crystallizes in a face-centered cubic structure with $a_0=6.083$A, and a small percentage of $\alpha-$Be.

SUMMARY

The results obtained here showed that beryllium crystallizes in two hexagonal structures. The lattice parameters of the α-form are: $a_0=2.286$A, $C_0=3.588$A, and $C_0/a_0=1.570$, and of the β-form, $a_0=6.93$A, $C_0=11.35$A, and $C_0/a_0=1.638$. The volume of the unit cell of the β-form is about 29 times of the α-form. These parameters compare favorably with those reported by other investigators. The amount of crystalline material in the β-form appears to be very small and hence can be confused with impurities.

The limit of solubility of gold in beryllium is between 2 and 3 atomic percent, and practically all of it is dissolved in the β-form. Since this amount of gold in beryllium produces no observable effects in the α-form, but marked changes in the β-form, it would seem that the unit cell of one form is not related to that of the other and the two forms crystallize independent of each other.

Most of the prominent lines of the β-form, as seen in Table I, appear at relatively small Bragg angles; therefore, they are beyond the range of measurements made with the back-reflection method.[4]

The technique of employing different alloying properties of allotropic forms of metallic elements to prove their existence seems to work very well as demonstrated in this investigation.

The cooperation of Dr. Frank Foote and Mr. James Schumar of Argonne National Laboratory for getting alloys made for us is gratefully acknowledged.

Single Wave-Length X-Rays for Powder Diffraction

L. K. Frevel and P. P. North
The Dow Chemical Company, Midland, Michigan
(Received May 10, 1950)

WL_α radiation filtered through a Cu-Zn foil is essentially monochromatic and has the advantage over filtered CuK_α or MoK_α radiation in that the $WL_{\alpha 1}$ line is approximately ten times as intense as the $WL_{\alpha 2}$ line. Moreover the comparatively wide $\Delta\lambda$-spread ($WL_{\alpha 2}-WL_{\alpha 1}$) results in very sharp powder lines at those Bragg angles where the CuK_α doublet or MoK_α doublet is unresolved.

THE geometrical factors deciding the resolution of a Debye-Scherrer-Hull pattern are: R, radius of the camera; r, radius of the powder specimen; h, length of the specimen; δ, divergence of the impinging x-ray beam; and σ, difference in wave-length of the K_α doublet. For a fixed camera radius the optimum resolution corresponds to the limiting condition: $r, h, \delta, \sigma \to 0$. The geometrical resolution for a camera of radius R can be expressed approximately as the sum of δ, the divergence of the diffracted beam due to the divergence or convergence of the impinging primary beam, and $\omega(r, h)$, the angular spread of the diffracted beam due to the dimensions of the powder specimen [$\omega(r, 0) = 2r/R$ radian, for parallel x-rays].[1-3] Accordingly, two powder reflections d and $(d+\Delta d)$ will be resolved if

$$(\theta_{\alpha 2}-\theta_{\alpha 1})_d < (\theta_d - \theta_{d+\Delta d})_\alpha > (\delta+\omega)_\theta,$$

or

$$(\delta+\omega)_\theta < |(\theta_{d+\Delta d})_{\alpha 1} - (\theta_d)_{\alpha 2}|$$

where θ is the Bragg angle for the interplanar spacing d, and α denotes the wave-length under consideration. For example, with $MoK_{\alpha 1}$ radiation the two powder reflections corresponding to 2.260A and 2.274A will be separated by 3.4′, which interval is equal to the MoK_α doublet separation, $(\theta_{\alpha 2}-\theta_{\alpha 1})_{2.26}$. Evidently the use of very narrow slits does not aid greatly in the resolution of a non-cubic powder pattern because of the disturbing doublet separation. The utility and precision of powder diffraction work would be greatly enhanced if an intense source of strictly monochromatic $L_{\alpha 1}$ radiation could be made available (see Table I). Crystal-monochromatized $L_{\alpha 1}$ radiation should be quite excellent for single crystal diffraction techniques; however, for powder diffraction work such a source of monochromatic x-rays would be too weak for most analytical applications. To circumvent this limitation it was decided to use filtered L_α radiation. An examination of Table I reveals that tungsten offers the most convenient source of L_α radiation because of the general availability of x-ray tubes with tungsten targets and because of the absence of target contamination by sputtering or evaporation from

Table I. The L-series spectra and the K-absorption edges of filters for L_β and L_γ.

Z	$\alpha 1$	$\alpha 2$	$\beta 1$	$\beta 2$	$\gamma 1$	Z_f	K_∞
57 La	2.665A	2.674A	2.458A	2.302A	2.141A	22 Ti	2.496A
58 Ce	2.561	2.570	2.356	2.208	2.048	22 Ti	2.496
59 Pr	2.463	2.473	2.259	2.119	1.961	23 V	2.267
60 Nd	2.370	2.381	2.166	2.035	1.878	23 V	2.267
62 Sm	2.199	2.210	1.998	1.882	1.726	24 Cr	2.070
63 Eu	2.120	2.131	1.920	1.812	1.657	24 Cr	2.070
64 Gd	2.046	2.057	1.847	1.746	1.592	25 Mn	1.896
65 Tb	1.976	1.986	1.777	1.682	1.530	25 Mn	1.896
66 Dy	1.909	1.920	1.710	1.623	1.473	25 Mn 26 Fe	1.896 1.743
67 Ho	1.845	1.856	1.647	1.567	1.417	26 Fe	1.743
68 Er	1.785	1.796	1.587	1.514	1.364	26 Fe 27 Co	1.743 1.607
69 Tm	1.726	1.738	1.530	1.463	1.316	27 Co	1.607
70 Yb	1.671	1.682	1.476	1.416	1.268	27 Co 28 Ni	1.607 1.487
71 Lu	1.619	1.630	1.424	1.370	1.222	27 Co 28 Ni	1.607 1.487
72 Hf	1.569	1.580	1.374	1.327	1.179	28 Ni 29 Cu	1.487 1.380
73 Ta	1.522	1.533	1.327	1.285	1.138	28 Ni 29 Cu	1.487 1.380

Table I.—(*Cont.*)

Z	$\alpha 1$	$\alpha 2$	$\beta 1$	$\beta 2$	$\gamma 1$	Z_f	K_∞
74 W	1.476	1.487	1.282	1.245	1.098	29 Cu 30 Zn	1.380 1.284
75 Re	1.433	1.444	1.239	1.206	1.061	29 Cu 30 Zn	1.380 1.284
76 Os	1.392	1.402	1.197	1.171	1.025	29 Cu 30 Zn	1.380 1.284
77 Ir	1.351	1.363	1.153	1.135	0.991	30 Zn 31 Ga	1.284 1.192
78 Pt	1.313	1.325	1.120	1.102	0.958	30 Zn 31 Ga	1.284 1.192
79 Au	1.277	1.288	1.083	1.070	0.927	31 Ga 32 Ge	1.192 1.118
80 Hg	1.242	1.253	1.049	1.040	0.897	31 Ga 32 Ge	1.192 1.118
81 Tl	1.207	1.218	1.015	1.010	0.868	31 Ga 32 Ge 33 As	1.192 1.118 1.045
82 Pb	1.175	1.186	0.983	0.983	0.840	32 Ge 33 As	1.118 1.045
83 Bi	1.144	1.155	0.952	0.955	0.813	32 Ge 33 As 34 Se	1.118 1.045 0.980
90 Th	0.956	0.968	0.765	0.794	0.653	35 Br 36 Kr 37 Rb	0.920 0.866 0.816
92 U	0.911	0.922	0.720	0.755	0.615	36 Kr 37 Rb 38 Sr	0.866 0.816 0.770

[1] M. J. Buerger, *X-Ray Crystallography* (John Wiley and Sons, Inc., New York, 1942), pp. 397–434.
[2] *International Tables for Determination of Crystal Structures* (Gebrüder Borntträger, Berlin, 1935), Vol. II, pp. 581–4.
[3] H. Lipson and A. J. C. Wilson, J. Sci. Inst. 18, 144 (1941).

TABLE II. Comparison of powder diffraction patterns of $BaSO_4$.

MoKα	CuKα	$WL_{\alpha 1}$	$WL_{\alpha 2}$	{hkl}
d	d	d	d	
4.444A	4.418A	— A	A	200
4.332	4.324	—		011
3.893	3.905	3.917		111
3.787	3.775			201
3.580	3.572	3.586		002
3.443	3.449	3.452		210
3.316	3.320	3.322	3.322	102
3.104	3.104	3.106		211
2.836	2.835	2.836		112
2.725	2.729	2.730		020
2.485	2.481	2.482		212
2.322	2.326	2.321		220
		2.302		103
	2.280	2.280		302
2.209	2.214	2.207		221
	2.123	2.120		113
	2.110	2.104	2.102	312, 122
	2.058	2.054		410
	1.926	1.927		321
	1.860	1.859		303
	1.791	1.789		004
	1.760	1.758		031, 313, 123
	1.728	1.727		501
		1.681*		230
	1.678	1.679		421
	1.640	1.639		231
	1.594	1.596		502
	1.534	1.536		323
		1.528*		304, 512
	1.494	1.4954		024
	1.477	1.4751		314, 124
	1.459	1.4567		521
		1.4275*		503
	1.425	1.4234		332
	1.405	1.4065		430
		1.4020*		611
	1.383	1.3839		015
		1.3798*		431, 513
	1.365	1.3624		040, 205
	1.352	1.3492		414
		1.3260*		141
	1.323	1.3206		215
	1.301	1.2996		240, 333
	1.263	1.2619		523, 504
		1.2527		315, 125
		1.2267		514
		1.2186		711, 341, 225
		1.2027		405
		1.1953		532
		1.1472		630, 441

Radiation:	MoKα = 0.7107A	CuKα = 1.5418A	$WL_{\alpha 1}$ = 1.4763A
Camera radius:	207.3 mm	71.7 mm	88.8 mm
Slit dimensions:	0.15 × 5 × 50 mm	0.4 mm rad. × 70 mm	0.1 × 3 × 50 mm
Specimen radius:	0.2 mm	Wedge method	0.4 mm

For very absorbing materials such as $BaSO_4$, the wedge technique with CuKα gives very sharp lines because only the topmost part of the wedge is effective for diffraction. Pyrex capillary tubing was used to hold the powder sample for the MoKα pattern, and polystyrene capillary for $WL_{\alpha 1}$. Since the sample radius was excessively large in the latter case, only the top-half of the sample was bathed in the x-ray beam, this arrangement being comparable to the wedge technique.

* Resolved reflections.

the tungsten filament. A brass filter, 0.025 mm thick, was used to absorb the WL_β and WL_γ lines. The WL_α doublet thus transmitted has a decided advantage over filtered K_α doublets because the $WL_{\alpha 2}$ line is only one-tenth as intense as the $WL_{\alpha 1}$ line; so that for all practical purposes powder patterns taken with this filtered radiation will correspond to patterns taken with $WL_{\alpha 1}$. The comparatively wide $\Delta\lambda$-spread ($WL_{\alpha 2} - WL_{\alpha 1} = 0.0111$A) has the further advantage of registering very sharp powder lines at those Bragg angles where the CuKα doublet or MoKα doublet is unresolved.

TABLE III. Powder diffraction data of Ag_2HgI_4.

CuKα d	CuKα I/I$_1$	$WL_{\alpha 1}$ d	$WL_{\alpha 1}$ I/I$_1$	$WL_{\alpha 2}$ d	$WL_{\alpha 2}$ I/I$_1$	{hkl}
6.273A	0.02	— A	—	A		200
5.635	0.03	—	—			210*
4.474	0.03	—	—			220
		4.198	0.03			300, 221*
		3.871	0.02			Phase II
3.646	1.00	3.647	1.00			222
3.502	0.02	3.497	0.10			320*
2.823	0.03	2.825	0.02			420
2.756	0.04	2.763	0.01			421*
		2.645	0.01			Phase II
2.570	0.02	2.576	0.10			422
2.475	0.01					510, 431*
		2.366	0.03			520, 432*
2.343	0.02	2.348	0.03			Phase II
2.231	0.50	2.233	0.70	2.234	0.01	440
		2.191	0.01			522, 441*
		2.137	0.30			531*
2.097	0.02	2.102	0.01			600, 442
2.076	0.02	2.079	0.01			610*
		2.015	0.01			Phase II
1.999	0.01	1.999	0.01			620
1.904	0.30	1.904	0.70			622
1.883	0.01					630, 543*
		1.823	0.10			444
1.732	0.01					720, 641*
1.686	0.02	1.686	0.01			642
1.616	0.02	1.617	0.02			650, 643*
1.576	0.04	1.578	0.06			800
1.530	0.01	1.532	0.01			820, 644
1.521	0.01	1.521	0.01			821, 742*
		1.509	0.01			653*
1.485	0.01	1.486	0.01			822, 660
1.449	0.10	1.449	0.30			662
1.440	0.01	1.440	0.02			832, 654*
		1.386	0.02			911, 753*
1.378	0.01	1.380	0.01			842
1.347	0.01					664
		1.367	0.01			920, 760*
1.310	0.01					852*
1.288	0.10	1.289	0.30			844
1.257	0.01					10·1·0, 942, 861, 742*
1.236	0.01	1.234	0.02			10·2·0, 862
1.214	0.04	1.216	0.03			10·2·2, 666

Legend: Filtered CuKα and WL_α x-rays were used to obtain the powder diffraction patterns of Ag_2HgI_4. {hkl} are the indices of the various reflections based on the unit cell $a = c = 12.62 \pm 0.02$A. The complex iodide Ag_2HgI_4 was prepared by the dropwise addition of $AgNO_3$ solution to an agitated solution of K_2HgI_4. Chemical analysis of the washed and vacuum-dried precipitate showed 23.51 percent Ag (23.35 percent theor.), 53.25 percent I (54.94 percent theor.), and 20.52 percent Hg (21.71 percent theor.). The sample for the WL_α exposure had been aged for five years in a closed vial.

* Reflections requiring larger unit cell than $a = c = 6.35$A.

Table II compares the powder patterns of $BaSO_4$ obtained with filtered CuKα, MoKα, and WL_α. It will be noted that six reflections were resolved with $WL_{\alpha 1}$ that appeared as unresolved lines with CuKα. Only two reflections out of fifty were attributable to $WL_{\alpha 2}$. For indexing non-cubic substances or reducing the superposition of powder lines in complex mixtures the use of $WL_{\alpha 1}$ radiation is to be recommended. To minimize continuous background radiation the tungsten x-ray tube with line focus should be operated at ~20 kv and ~40 ma.

The complex iodide, Ag_2HgI_4, is reported by Ketelaar[4] to be tetragonal with $a = c = 6.34kX$. Patterns were taken of this substance with CuKα and $WL_{\alpha 1}$ radiation to ascertain if any splitting could be detected for pyramid reflections. Whereas no splitting was revealed (see Table III), a sufficiently large number of weak reflections were observed to require a larger unit cell. Extensive grinding of the complex iodide diffuses these additional sharp reflections into faint broad bands. It is quite likely that one is dealing with an order-disorder system in Ag_2HgI_4 and that the $E3_1$ structure will have to be reexamined in the light of the present findings.

[4] J. A. A. Ketelaar, Zeits. f. Krist. 80, 190–203 (1931).

Shielding Properties of the Concrete Wall of the M.I.T. Cyclotron

VICTOR DELANO* AND CLARK GOODMAN
Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received May 15, 1950)

The shielding properties of the four-foot concrete wall of the M.I.T. cyclotron have been studied, using foils for neutrons and x-ray films for gammas. Fast neutrons were measured by counting the beta-active products of $Al^{27}(n, p)Mg^{27}$, $Fe^{56}(n, p)Mn^{56}$, and $Al^{27}(n, \alpha)Na^{24}$ reactions, thermal activation being eliminated by Cd holders. Resonance (1.44 ev) and thermal neutron distributions, measured by Cd-In-Cd and In, with suitable corrections, were reduced to absolute intensities by comparison with a known flux. Measurements were made with cylindrical concrete blocks fitted into a 4½" diameter hole which extended through the concrete wall. The results indicate an attenuation factor of 8.3 per foot for the fast neutron component and 10 per foot for thermal neutrons. An estimate of the fast neutron flux above 4.5 Mev showed a transmitted intensity slightly higher than the presently accepted tolerance value when bombarding Be. The absolute thermal neutron flux was below the tolerance level. The r.m.s. distance and the diffusion length for thermal and resonance neutrons in concrete were estimated by considering the distributed source to be equivalent to a point source at the inner face of the wall. The gamma-ray measurements were not convertible to roentgens. A mean energy of 2.7 Mev was obtained for the gamma-radiation within the wall. The attenuation factor for total gamma-radiation in the wall at large distances from the inner face was 5.9 per foot.

INTRODUCTION

THE occurrence of radiation cataract among the Japanese survivors at Hiroshima and Nagasaki and among cyclotron workers in this country has emphasized the importance of adequate neutron shielding. The problem is complicated by the invariable presence of gamma-rays which arise both from the reactions producing the neutrons and from the absorption and scattering of neutrons within the shield. The latter effect imposes important limitations on the choice of shielding material. What is needed is a combination of heavy and light atoms which are effective for attenuating gamma-rays and neutrons respectively. Of the many materials and combinations of materials investigated in the search for an ideal shield, concrete is one of the most satisfactory especially when cost is a major consideration and where size and bulk are not of too great concern.

While concrete has found wide use for the shielding of particle accelerators, there is only a limited amount of published information regarding its effectiveness.[1] Since the radiations from the M.I.T. cyclotron are representative of a number of existing sources and of others being built, the results of the tests made on this four-foot concrete shield may be of general interest.

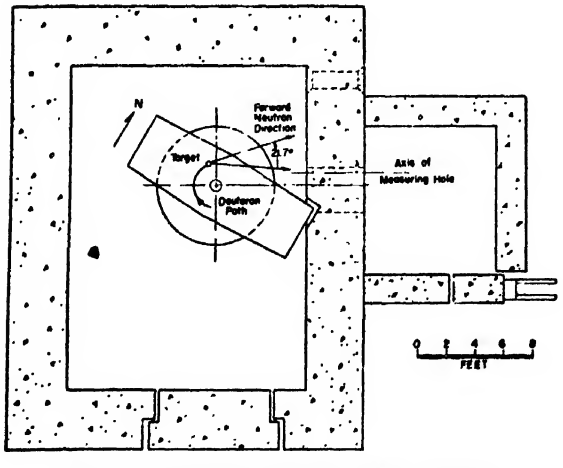


FIG. 1. Plan view of cyclotron shield, showing location of measuring hole.

EXPERIMENTAL ARRANGEMENTS AND PROCEDURES

Description of Cyclotron Shield and Measuring Hole

The M.I.T. cyclotron is shielded by a four-foot concrete wall topped with a three-foot concrete roof. Through the northeast wall of the shield, of which a plan view is shown in Fig. 1, a rectangular opening three feet wide and one foot high is provided for emergent beam experiments. When this beam is not in use, the opening can be closed with an array of concrete blocks of three sizes: 6"×6"×11½", 3"×6"×11½" and 6"×6"×6". These sizes are approximate because of slight surface irregularities.

Because of these irregularities and the relatively large dimensions, another set of blocks was constructed which formed a smooth, horizontal hole through the wall into which one could slide close-fitting cylindrical blocks for intensity measurements with foils and films. Since the center of the rectangular opening is below the central plane of the cyclotron dees, these blocks were designed to fit in the upper left-hand corner of the rectangular opening (see Fig. 2). By this arrangement the measuring foils could be exposed to the most energetic portion of the neutron beam.

As seen in Fig. 1, the axis of the measuring hole was displaced from the direction of the deuteron beam at the cyclotron target. For 15-Mev deuterons on Be, the maximum energy of neutrons to which the measuring hole was exposed was calculated as 18.8 Mev.

* Lieutenant Commander, U. S. Navy, now at Los Alamos Scientific Laboratory.

[1] M. S. Livingston, J. App. Phys. 15, 2–19, 128–147 (1944); E. Creutz and K. Downes, J. App. Phys. 20, 1236–1240 (1949); P. C. Gugelot and M. G. White, Phys. Rev. 74, 1215 (1948); P. Jensen and O. Ritter, Zeits. f. Naturforschung 2a, 376–384 (1947).

As shown in Fig. 3, three *C* blocks formed a base along the bottom of the rectangular opening for the *A* and *B* blocks. The measuring hole was formed using two pairs of *A* blocks, placed end to end. Two *B* blocks were set alongside these (see Fig. 2). All blocks were arranged to eliminate, as completely as possible, any continuous cracks or slits through the opening. Because of the small dimensional discrepancies in the latter, it was necessary to use wooden wedges between the blocks and the left-hand edge of the rectangular opening in order to force the hole-forming blocks tightly against the cyclotron wall. The original rectangular blocks were installed with their surfaces flush with the cyclotron wall in the cyclotron vault but they failed to complete the rectangular opening by about three-quarters of an inch along the outer side of the wall.

Each cylindrical *D* block has a $\frac{3}{16}''$ recess in one end in which were placed foils for neutron measurements or films for gamma-radiation measurements. By using two different lengths (2″ or 4″), measurements could be taken at two-inch intervals through the wall. "*A*" runs, as they are henceforth called, used twelve four-inch blocks. "*B*" runs used one two-inch block followed by eleven four-inch blocks and, at the end of the hole, another two-inch block. All the four-inch blocks were numbered and the two-inch blocks lettered to insure that successive "*A*" and "*B*" runs would have the same sequence of blocks.

The cylindrical blocks were always inserted with their recessed ends toward the cyclotron. The blocks were inserted by hand and removed with piano wire attached in a manner allowing half the blocks to be removed from the front of the hole and half from the back. Clearance between the top of the cylindrical blocks and the roof of the hole was $\frac{1}{16}''$ or less. Flake graphite was used as a lubricant.

Composition of Concrete

All the blocks used in the rectangular opening of the cyclotron wall were made from a concrete mix consisting of 1 part cement, $1\frac{1}{2}$ parts sand and 2 parts gravel (pea stone, $\frac{1}{2}''$ gravel), with about 5–$5\frac{1}{2}$ gallons of water per 100 lb. bag of cement. Composition of the cyclotron wall was slightly different, being made from a mix of 1 part cement, 2 parts sand and 4 parts gravel (1″ grated to $\frac{1}{4}''$), by weight. The difference in the mixes was necessitated by the relatively small size of the blocks which required more cement to insure adequate strength and proper shape. Table I gives chemical analyses of the two cements. Table I also includes the assumed composition of the cement used in the work of Jensen and Ritter.[2] The composition of the sand and gravel given in Table I is based on typical analyses of the sand and gravel from glacially deposited pits near Scituate, Massachusetts. Jensen and Ritter considered their sand and gravel combination to be pure SiO_2.

[2] P. Jensen and O. Ritter, see reference 1.

FIG. 2. Photograph of measuring hole showing insertion of foil and film-holding blocks.

An analysis of a sample block very kindly made by the Oak Ridge National Laboratory gave an average water content of 7.6 percent by weight. Using this value we have calculated the weight percentages of elements present as given in Table II.

The average density of the concrete in the cyclotron wall was measured as 2.20 g/cc, slightly less than Jensen and Ritter's value of 2.26 g/cc.

Counters and Counting Circuits

The beta-activities of the foils were measured with a Victoreen, mica-window, counter (Model VG-Special) mounted in a $1\frac{1}{2}''$ lead shield. The end-window of the counter was 2.0 mg/cm^2; the foils were inserted on brass trays with a $\frac{1}{8}''$ deep recess, 3.18 cm in diameter centered directly below the counter window. The distance from the bottom of the tray to the center of the window was 1.0 cm. Circuitry consisted of a one-stage amplifier, a discriminator stage with adjustable bias and a scale-of-64, followed by a mechanical register.

For standardizing the counting rates, a brass slide was provided containing a small amount of RaD+E covered with 6 mils of aluminum which allowed only the 1.17-Mev beta-rays of RaE to emerge. The counter was shown to be linear up to at least 20,000 counts per

TABLE I. Composition of concrete mix used in cyclotron wall.

	Cement			Sand and gravel	
	Cambridge Cement Stone Company (movable blocks)	Boston Sand and Gravel Company	Cement used by Jensen and Ritter	Sand	Gravel
CaO	63.41%	63.3% (est.)	65%	0.25%	1.08%
SiO_2	20.03	22.2	25	95.87	73.60
Al_2O_3	7.10	5.1	7	1.83	14.44
Fe_2O_3	3.14	4.9	3	0.04	0.43
MgO	2.84	1.4			
SO_3	1.88	1.5			
FeO				0.27	1.49
Na_2O				0.88	4.20
K_2O				0.61	4.46
H_2O				trace	trace
Misc.	1.60	1.6			

minute. Active foils were allowed to decay below 15,000 counts per minute, although less than 10,000 counts per minute was generally used. The average of backgrounds taken before and after each set of measurements was used, and counting rates less than twice the average background were discarded.

Fast Neutrons—Aluminum and Iron Foils

As calculated from Bethe's penetrability function,[3] the (n, p) reaction reaches half its maximum cross section for Al and Fe at 4.5 and 7.6 Mev, respectively. Al also has an (n, α) reaction reaching half the maximum cross section at 8.3 Mev. A single set of Al foils gives results for both energy intervals. Reynolds 24-mil (165 mg/cm^2) Al sheets, with a purity of 99.8 percent, were die-cut into disks $1\frac{3}{16}''$ in diameter. Fe foils were die-cut from Armco Magnetic Ingot Iron into disks of the same diameter from sheets hand-rolled to 10 mil (200 mg/cm^2). After activation, the Al foils exhibited only the 10.2 min. half-life from $Al^{27}(n, p)Mg^{27}$ and the 14.8 hr. half-life[4] from $Al^{27}(n, \alpha)Na^{24}$, indicating that any impurities were negligible. The 2.59 hr. activity[5] from $Fe^{56}(n, p)Mn^{56}$ was observed but in addition there was an apparent half-life of about 28 hr. A correction was made for the 28 hr. period.

To reduce the effect of (n, γ) reactions, the Al and Fe foils were encased in 1 mm thick Cd holders. In the Fe foils, no (n, γ) activation was detectable. By waiting 20 min. before counting, the 2.4 min. activity in Al became negligible.

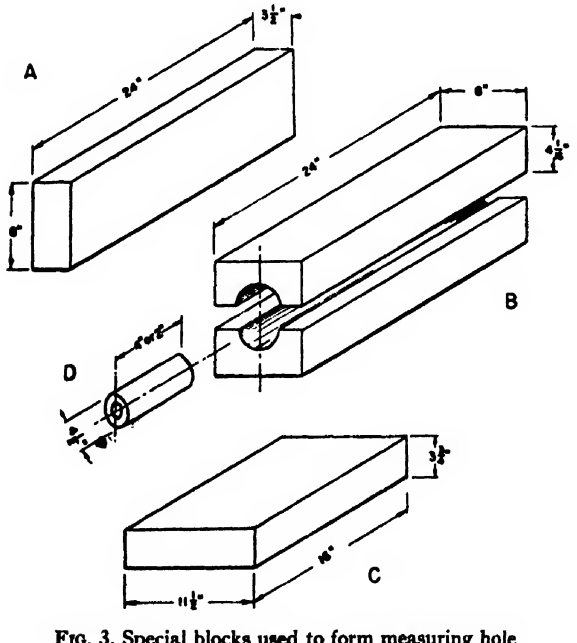


FIG. 3. Special blocks used to form measuring hole and to hold foils and film.

[3] H. A. Bethe, Rev. Mod. Phys. 9, 69 (1937).
[4] K. Way, *et al.*, Nucleonics 2 (No. 5), 104 (1948).
[5] Segrè Chart, revised April 1948 (AECD-2111).

TABLE II. Average composition of concrete.

H	0.85
O	50.64
Si	30.49
Al	4.44
Ca	8.03
Fe	1.19
Mg	0.23
Na	1.66
K	1.87
S	0.12

Thermal and Resonance Neutrons—Indium Foils

Indium foils were used for measuring thermal and resonance neutron distributions by the Cd-difference method. Foils 1 cm in diameter were cut from 99.97 percent in sheets hand-rolled to a thickness of 3 mil (56 mg/cm^2). Activation indicated impurities were negligible. Because the In foils were bombarded only short lengths of time, the 4.5 hr. activity was relatively weak, and correction was made only when it was significantly above background. In foils were counted front and back after bombardment, since the sum of the front and back activity is proportional to the flux for thermal and 1.44-ev resonance neutrons.

Indium Foil Calibration and Corrections

Attempts were made to calibrate the indium foils with the cyclotron as a source, but, owing to nonhomogeneity of the neutron distribution in the cyclotron vault, this was unsuccessful. Instead, a 1-gram Ra-Be source in a 4-foot cubical water tank was used. In addition to corrections for variations in individual foils, two other factors were considered before the foil activities could be interpreted in terms of absolute neutron flux. These factors were (1) the attenuation of 1.44-ev resonance neutrons by the 1-mm Cd holder used for resonance measurements and (2) the distortion of the thermal neutron flux by the detector. Both factors are thoroughly discussed by Bothe[6] and his results are adapted to this work. Bothe develops an expression for the absorption of a neutron striking a detector in any direction:

$$\alpha = 1 - e^{-\mu\delta}(1 - \mu\delta) + \mu^2\delta^2 E_i(-\mu\delta)$$

where α = average absorption probability for a neutron striking in any direction, δ = linear thickness of absorbing material, $\mu = N\sigma$, and E_i = exponential integral.

With this absorption factor, the correction factor to be applied to all resonance runs (Cd holders) becomes:

$$f_a = 1/(1 - \alpha).$$

Inserting the appropriate values for Cd, $\delta = 0.1$ cm, $N = 0.0463 \times 10^{24}$ nuclei/cc,[7] and $\sigma_{1.44} = 11.5$ barns,[8] we obtain $f_a = 1.11$.

[6] W. Bothe, Zeits. f. Physik 120, 437–449 (1943).
[7] H. R. Kroeger, Nucleonics 5 (No. 4), 51–54 (1949).
[8] Goldsmith, Ibser, and Feld, *The Science and Engineering of Nuclear Power*, Vol. I, p. 456 (1947).

To correct for distortion, the depression factor for foils which are small compared with the scattering path length in the medium was taken from Bothe's work using λ_{tr} instead of λ_s:[9]

$$f=\frac{1}{1+0.23\alpha R_K/\lambda_{tr}}$$

where f=depression factor, α=absorption factor, previously described, R_K=radius of foil, λ_{tr}=transport mean free path.

Inserting appropriate values for In, $\alpha=0.100(\sigma_{th}=190$ barns), $R_K=0.5$ cm, $\lambda_{tr}=1.95$ cm, we obtain $f=1/1.00590$, or the correction factor equals 1.00590.

After applying these factors to the resonance and Cd difference results, the corrected, saturated initial activities were converted to absolute neutron flux.

Gamma-Ray Measurements

Measurements of total gamma-radiation proved considerably more difficult than neutron measurements. X-ray films (duPont 552) were used. This film is paper-wrapped in dental-size packets, each containing two films, one sensitive and one insensitive, such that 0.02 to 20 roentgens can be measured. The exposed films are examined with an Ansco-Sweet Densitometer and the density is converted to dosage by means of calibration charts. These charts are based on a 1-mm layer of Cd surrounding the film.

The film has two major undesirable characteristics. The bare film is severely blackened by beta-rays and electrons. Furthermore, the blackening is constant for gamma-radiation above 300 kev. Below 300 kev, the blackening per roentgen rises sharply. When only alpha-, beta- and gamma-radiations are present, 1 mm of Cd is sufficient to attenuate the charged radiations and has been shown to be adequate for reducing the blackening from the soft gamma-radiation. When there is a high thermal neutron flux, as in most of the present measurements, the secondary gammas from radiative capture of neutrons in Cd produce serious extraneous blackening. Kikuchi, Aoki and Husimi[10] found that of 32 elements studied, Cd is outstanding in its production of capture gammas.

Summary of Runs

In order subsequently to refer to runs by number those providing useful data are listed in Table III.

RESULTS AND DISCUSSION

Fast Neutron Distribution

For the fast neutrons we assume a point source at the cyclotron target and an exponential attenuation through the wall, i.e., a $(1/r^2)\exp(-r/L_r)$ distribution,[11] where r is the distance of the foil from the target and L_r is the relaxation length at large distances.

Figure 4 shows one set of measurements of fast neutron attenuation through the concrete wall for Al and Fe detectors. We would expect the relaxation lengths for these curves to be the same for the three activities observed since the energies involved are nearly the same. At large distances in the wall, it would appear that the fast, resonance, and thermal neutrons should have the same relaxation length, since the higher energy feeds into the lower energies and the diffusion length of low energy neutrons is quite small. Table IV shows the results.

Values for both thresholds of the Al foils agree and are considered the most acceptable for the threshold range covered. The attenuation factor corresponding to a relaxation length of 14.4 cm is 8.3 per foot.

Dementi and Timoshuk's[12] cross section for the Al(n, p) reaction is 15 millibarns and the extrapolated short-lived activity at the wall's outer surface is 5.75 c.p.m. for Run 14. This run is used because its average beam current was the strongest of the Al runs, 164 μamp., and its relaxation length the same as the average, 14.4 cm.

The flux is estimated from the following relation:

$$(nv)_f=A/(\sigma_p NV),$$

where $(nv)_f$=flux in neutrons/cm^2/sec., A=absolute foil activity in disintegrations per second, σ_p=reaction cross section in cm^2, N=nuclei per foil volume, and V=foil volume.

This gives a flux of 243 neutrons/cm^2/sec. above the Al(n, p) threshold and gives an interesting comparison with the tolerance flux of 190 n/cm^2/sec. for 4.5-Mev neutrons determined from:[13]

$$(nv)_f=1.45\times10^{-21}/\sigma_H E,$$

TABLE III. Summary of runs.

Run no.	Detector	Type run	Length	Av. beam current (μ amp.)	Target	Remarks
1	Al	*A*	2 h	60	Be	
4	Al	*B*	20 m	96	Be	
9	Fe	*B*	11¼ h	155	Cu	
13	Al	*B*	34 m	150	Be	
14	Al	*A*	30 m	164	Be	
15	In	*A*	10 m	165	Be	Thermal+Resonance (Paper holders)
16	In	*A*	10 m	170	Be	Resonance (Cd holders)
17	Film	*A*	5 m	168	Be	
18	Film	*A*	10 m	170	Be	
19	Film	*B*	5 m	172	Be	
20	Film	*B*	10 m	168	Be	
21	Fe	*B*	6¼ h	165	Be	
22	Fe	*A*	7h-55 m	164	Be	
23	In	*B*	10 m	170	Be	Thermal+Resonance
24	In	*B*	10 m	162	Be	Resonance

[9] This significant correction was suggested by C. W. Tittle.

[10] Kikuchi, Aoki, and Husimi, Proc. Phys. Math. Soc. Japan 18, 115 (1936).

[11] R. F. Christy, MDDC-1175, p. 127.

[12] V. S. Dementi and D. V. Timoshuk, Comptes Rendus Acad. Sci. U.R.S.S. 27, 926 (1940).

[13] M. L. Goldberger, *The Science and Engineering of Nuclear Power* Vol. II, Chapter 12 (1949).

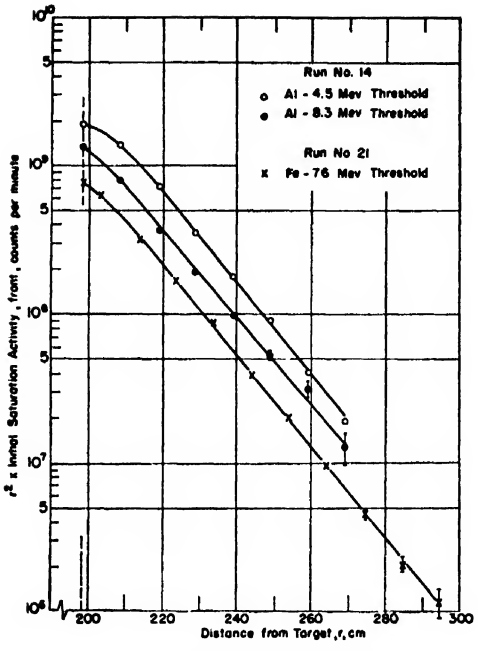


FIG. 4. Fast neutron distributions in concrete.

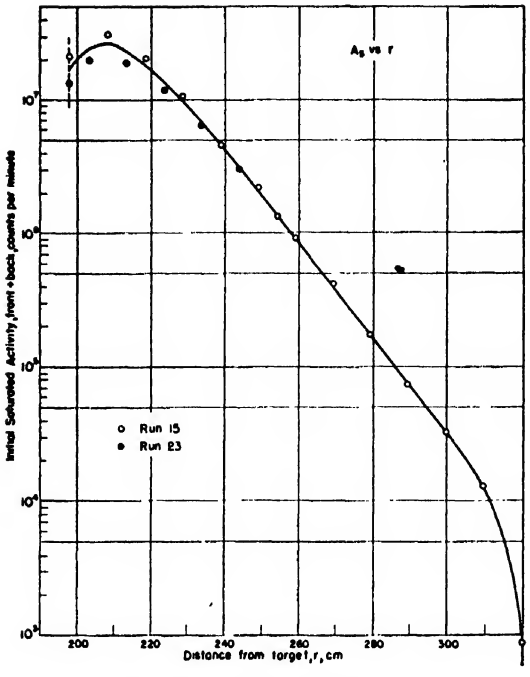


FIG. 5. Thermal plus 1.44 ev resonance neutron distributions in concrete.

where σ_H = scattering cross section for hydrogen and E = neutron energy in Mev. Our approximate value can also be compared with the tolerance flux of 85 n/cm²/sec. for 2–20 Mev neutrons reported by Mitchell.[14]

Our estimate is probably low, due to the decrease in reaction cross section as the Al(n, α) reaction begins to compete.

Thermal and Resonance Neutron Distributions

Since the cyclotron is non-reproducible, it was necessary to establish a relation between the resonance and thermal resonance distributions in order to provide a basis for normalization.

Table III shows that 15 and 23 were thermal+resonance runs and 16 and 24 were resonance runs. A foil was placed 5½″ to the right of the measuring hole axis on the inner face of the wall; on Runs 15 and 23, the foil was Cd covered, on Runs 16 and 24, paper covered. Between each of these specially placed foils and the regular foil nearest the cyclotron, thermal+resonance/resonance ratios were established. Since only one foil was covered with Cd on Run 23, with the depressive effect consequently less, the thermal+resonance/resonance ratio of this run (2.594) was selected to be applied to the mean values of the thermal+resonance and resonance runs in order to relate them.

Figure 5 shows a plot of the results of Runs 15 and 23, with 15 corrected to an average beam current of 170 μamp. by assuming the neutron yield to be linearly proportional to the beam current for small increments of current, together with the mean curve for the two runs. Figure 6 shows a plot of the results of Runs 16 and 24, with 24 corrected to 170 μamp. beam current and the resulting mean curve. The thermal+resonance/resonance ratio of 2.594 was applied to the initial point of the mean resonance curve of Fig. 6 to determine the starting point of the corresponding thermal+resonance curve. A point-by-point subtraction was carried out next and the foil depression correction applied to this result, giving the thermal distribution shown in Fig. 6.

TABLE IV. Relaxation lengths for Al and Fe threshold detector runs, from r^2A_s vs. r plots (in cm).*

Al (4.5 Mev)	Al (8.3 Mev)	Fe (7.6 Mev)
14.1 (1)	14.1 (1)	14.7 (9)
13.7 (4)	14.1 (4)	14.0 (21)
15.2 (13)	14.7 (13)	13.4 (22)
14.4 (14)	14.6 (14)	
Av. 14.4	14.4	14.0

* Numbers in parentheses indicate the number of the run.

[14] J. S. Mitchell, Brit. J. Radiology 20, 177 (1947).

Mean Square Distances and Diffusion Length

One of the first results obtainable from the thermal resonance data is the calculation of the mean square distance for thermal and resonance neutrons. Before reasonable values can be determined, however, an initial assumption is necessary for which no justification can be found in theory. The primary source of neutrons is naturally the cyclotron target, and at first glance it seems correct to consider a distribution based on the actual distance of each foil from the target, following

the usual $(1/r^2)\exp(-r/L_r)$ attenuation, where r is the distance from the target and L_r the relaxation length. Figure 7 shows such a plot. Amaldi and Fermi's equation can then be used to calculate mean square distance:[15]

$$\bar{r}^2=\frac{\int_0^\infty r^4 A dr}{\int_0^\infty r^2 A dr},$$

where r is the distance of the foil from the target and A is the foil activity as a function of distance.

Values determined by this method for the mean square distance for thermal and resonance neutrons are absurdly high, 5×10^4 cm^2, and have no significance, being obviously overweighted by the effect of the 200-cm distance in air between the target and the front face of the wall. By assuming that the thermal and resonance neutron source is located at the inner face of the wall, but continuing to treat it as a point source (as plotted in Fig. 8), reasonable values are obtained:

$$\bar{r}_{th}^2=2148 \text{ cm}^2 \text{ and } \bar{r}_{res}^2=1818 \text{ cm}^2.$$

The root mean square distance for the resonance neutrons, $\bar{r}_{res}=42.7$ cm, compares closely with Jensen and Ritter's value of 44.9 cm,[16] which is corrected for boundary disturbance and also is based on the use of a Ra-Be neutron source. Our result suggests a lower mean energy for the neutron spectrum of the cyclotron than for a Ra-Be source.

Continuing, we estimate the diffusion length L from the relation:[17]

$$\bar{r}^2_{thermal}=\bar{r}^2_{thermal\ source}+6L^2.$$

Although $\bar{r}^2_{thermal\ source}$ is required by the equation, we can accept the value of $\bar{r}^2_{res}$ as a close approximation. From this we obtain a diffusion length of 7.4 cm.

We can compare this with the calculated results: $L^2=\lambda_{tr}\lambda_a/3$, where λ_{tr} and λ_a are the transport and absorption mean free paths, respectively. Table V lists the cross sections used in computing λ_a, λ_s and λ_{tr}. The Mg and S in the concrete were neglected.

Table VI lists the calculated values of absorption, scattering and transport mean free paths, and the diffusion length. The value of the latter, 7.7 cm, is in good agreement with the experimental observed value of 7.4 cm given in Table VI.

Relaxation Lengths for Thermal and Resonance Neutrons

Assuming a thermal and resonance source at the inner face of the wall, we obtain relaxation lengths of

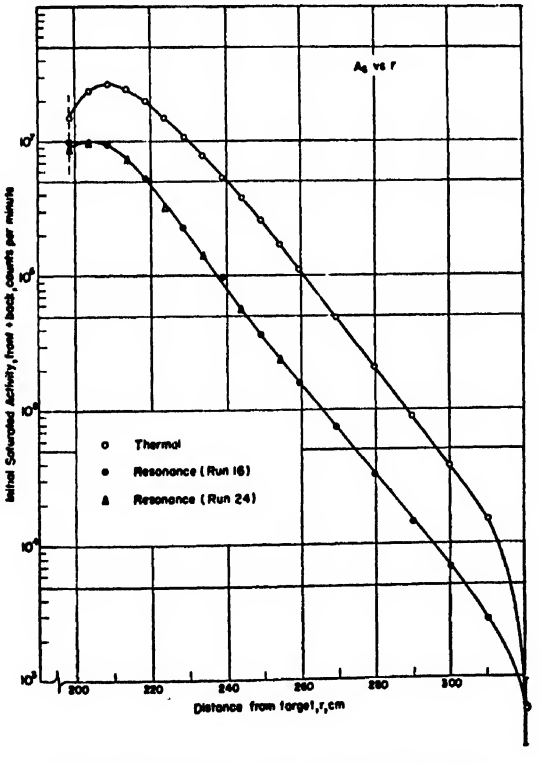


FIG. 6. Thermal and 1.44 ev resonance neutron distributions in concrete.

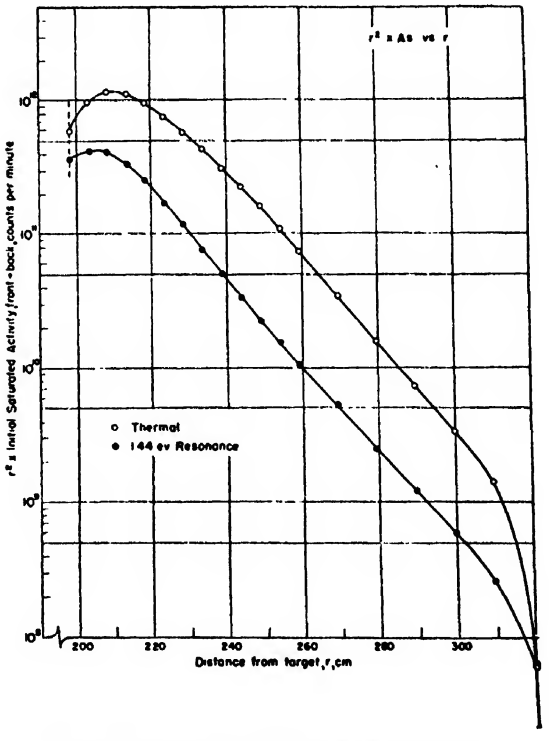


FIG. 7. Thermal and 1.44-ev resonance neutron distributions in concrete.

[15] E. Amaldi and E. Fermi, Phys. Rev. 50, 899–928 (1936).

[16] P. Jensen and O. Ritter, Zeits. f. Naturforschg. 2a, 376–384 (1947).

[17] A. M. Munn and B. Pontecorvo, Can. J. Research 25A, 157 (1947).

$L_{\tau(th)} = 16.9$ cm and $L_{\tau(res)} = 18.0$ cm. Jensen and Ritter[18] embedded their source in the center of a concrete block and reported relaxation lengths of 20.4 cm and 24 cm for thermal and 1.44-ev resonance neutrons, respectively. It should be noted that they used a half-life for In of 57 minutes, which would account for some of the discrepancy between our measurements and theirs. Their experimental arrangement was close to ideal, making it unnecessary to adopt the assumption of the source at the surface of the wall.

In order to compare our work with recent observations by Gugelot and White,[19] we consider our source as the cyclotron target and the thermal and resonance relaxation lengths thus obtained are 13.1 and 14.0 cm, respectively. Gugelot and White, using Ag detectors, report a half-value thickness of 9.5 cm for ordinary concrete with 8.5 percent water content. This corresponds to a relaxation length of 13.7 cm.

The relaxation length for thermal neutrons, based on the cyclotron target as the source, is equivalent to an attenuation factor of 10 per foot for slow neutrons, distinctly larger than the 8.3 per foot obtained for fast neutrons. Since the former is based on the observed thermal distribution, which is considerably more uncertain, a value of 9 per foot is considered a reasonable average for future design purposes.

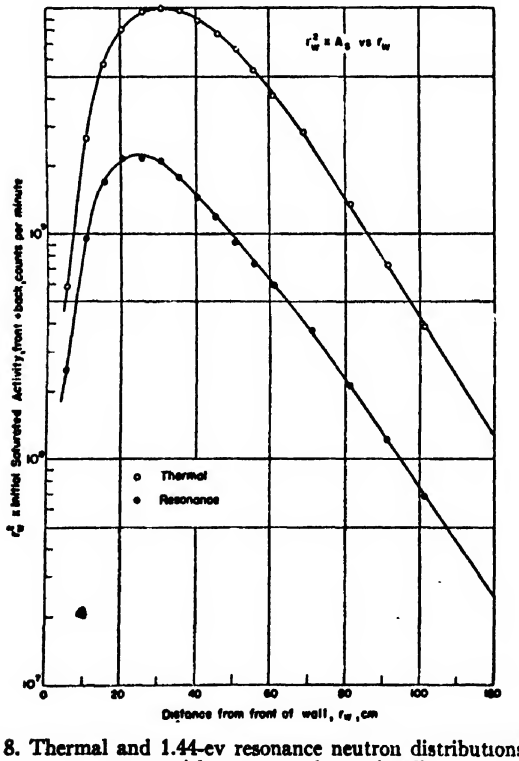


FIG. 8. Thermal and 1.44-ev resonance neutron distributions in concrete, with respect to front of wall.

Absolute Thermal and Resonance Neutron Fluxes

In the same four-foot water tank used for In foil calibration, Dacey and Paine[20] have accurately established the thermal and resonance flux at a point 14 cm from a 1-gram Ra-Be source. From the activities of In foils at this point, we obtain the following ratios for converting In foil activity from a cyclotron bombardment to neutron flux:

$$\frac{(nv)_{th}}{(A_s)_{th}} = \frac{8613}{6082} = 1.416,$$

$$\frac{(nv)_{res}}{(A_s)_{res}} = \frac{124.2}{483.5} = 0.2569.$$

These are then applied to the mean values for the thermal and resonance distributions, respectively, to obtain the thermal flux in neutrons/cm²/sec. and the resonance flux in neutrons/cm²/sec. per unit logarithmic energy interval.

It is interesting to compare the thermal flux at the outer surface of the wall with values calculated by Wende[21] and by Mitchell[22] as biologically equivalent to the *former* tolerance dosage of 0.1 *r*/8 hour day.[23] Considering only the effect on body tissue of gamma-rays from the $H'(n, \gamma)$ reaction, Wende estimates the tolerance flux to be 15,000 n/cm²/sec. for thermal neutrons. Mitchell, in a more realistic calculation, includes the

TABLE V. Thermal cross sections for nuclei in concrete.*

Nucleus	(Nuclei/cc) $\times 10^{24}$	σ_a	σ_s (in barns)	σ_{tr}
H	0.0112	0.32	60.4	26.6
O	0.0419	0.001	4.1	3.9
Si	0.0144	0.2	2.3	2.2
Al	0.0022	0.22	1.4	1.4
Ca	0.0027	0.5	3.5	3.4
Fe	0.00028	2.5	11	11
Na	0.00096	0.45	4	4
K	0.00063	2.5	1.5	1.5

* σ_s and σ_{tr} for H are from J. E. Dacey and R. W. Paine, Jr., S.M. Thesis, M.I.T. (1949). All other cross sections from H. R. Kroeger, Nucleonics 5 (No. 4), 51–54 (1949).

TABLE VI. Absorption, scattering and transport mean free paths and diffusion lengths (in cm).

λ_a	90.9
λ_s	1.11
λ_{tr}	1.95
L (calculated)	7.7
L (experimental)	7.4

[18] P. Jensen and O. Ritter, Zeits. f. Naturforschung **2a**, 376–384 (1947).

[19] P. C. Gugelot and M. G. White, Phys. Rev. **74**, 1215 (1948). Subsequent to the submission of the present manuscript, an article by these authors has appeared giving more complete data, J. App. Phys. **21**, 369 (1950).

[20] J. E. Dacey and R. W. Paine, Jr., S.M. Thesis, M.I.T. (1949).

[21] C. W. J. Wende, *The Computation of Radiation Hazards* Report No. M-1324 (TNX-7) (11 January 1944).

[22] J. S. Mitchell, Brit. J. Radiology **20**, 79 (1947).

[23] R. E. Lapp and H. L. Andrews, *Nuclear Radiation Physics* (Prentice-Hall, Inc., New York, 1948), Chapter 18.

effects of the $N^{14}(n, \gamma)$ and $N^{14}(n, p)$ reactions as well as the gammas from neutron capture by hydrogen and estimates the tolerance flux to be 12,500 n/cm²/sec. Our data indicate the thermal flux at the wall's outer surface to be 830±540 n/cm²/sec. It is evident, therefore, that for 170 μamp. average beam current and a Be target, the cyclotron shield has reduced the thermal flux more than an order of magnitude below either of these estimated tolerance levels and well below the recent more conservative level of 1500 n/cm²/sec.[24]

Gamma-Radiation

Figure 9 is typical of radiation measurements using bare film. The plot is in terms of $r^2 \times I$ as a function of r, where r is the distance in cm from the cyclotron target to the film and I the apparent ionization intensity in roentgens per hour. The values for I are determined as previously described. These intensities are somewhat uncertain because of the possible blackening caused by low energy gamma-rays (<300 kev) and by capture gammas from the cadmium cover.

Two prominent features of all the curves obtained are the increase in apparent ionization intensity in the first 10 cm of the wall because of secondary and capture gammas from neutron interaction within the wall and also the tendency toward linearity beyond 40–45 cm of wall distance.

At large distances in the shield three processes combine to produce the total gamma-radiation: (1) attenuation of primary gammas from the target (2–8 Mev); (2) capture gammas from thermal neutron capture by hydrogen (2.18 Mev); (3) secondary gammas from radioactive products of fast neutron reactions (0.5–2.8 Mev).[25] The relative weights of these effects cannot be determined, although (1) and (2) appear more important than (3).

From the curves, an average absorption coefficient of 0.058 cm^{-1} is obtained. Hirschfelder and Adams[26] have reported mass absorption coefficients for monoenergetic gamma-rays in concrete. Converting their results for 1- and 3-Mev gamma-rays into linear absorption coefficients, we obtain values of 0.0856 and 0.0517 cm^{-1}. Assuming the ionization intensity is directly proportional to the flux intensity over the energy range with which we are concerned, the absorption coefficient from our data suggests a "mean" gamma-ray energy of about 2.7 Mev. Gugelot and White,[27] using 18-Mev protons on Be, have reported a half-value thickness of 13.3 cm for gamma-rays in concrete of slightly higher density (2.35 g/cc) than used here. This corresponds to an absorption coefficient of 0.052 cm^{-1} and, according to Hirschfelder and Adams,[26] to a "mean" energy of about 3 Mev. The attenuation factor for the gamma-radiation in concrete, measured in the linear portion of the curves, is 5.9 per foot.

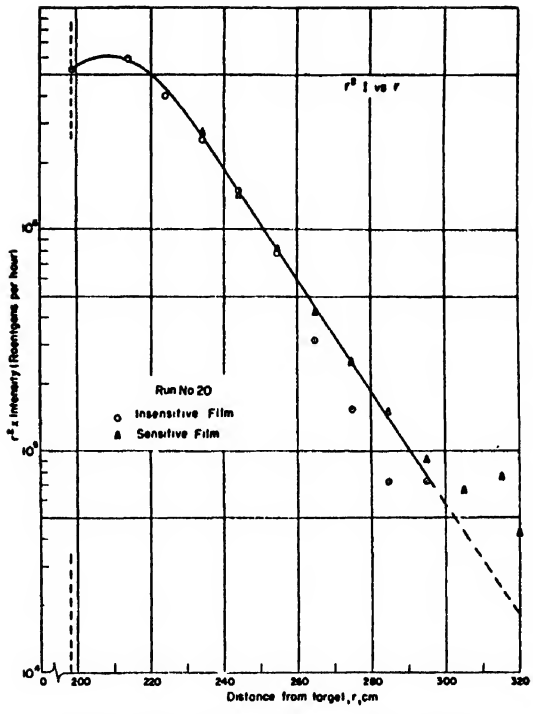


FIG. 9. Apparent ionization intensity distribution through concrete.

Average gamma-radiation from the cyclotron is probably about 5 Mev. The capture and secondary gammas, with lower average energies, appear sufficiently abundant to result in a lower mean energy for the total gamma-radiation in the wall. This seems to agree with Gugelot and White's[27] results.

This effect should be an important consideration in designing future cyclotron shields. With better information, the apparent importance of the capture and secondary effects may be verified. If so, cyclotron shield design should emphasize the importance of lower energy gamma-rays as much as those of higher energies produced at the target.

ACKNOWLEDGMENTS

The authors appreciate discussions of this work with Dr. M. Stanley Livingston and our colleagues in the Nuclear Shielding Group. The facilities of the cyclotron were made available by Dr. R. D. Evans with the assistance of Mr. J. B. Bulkley. The painstaking assistance of Miss Joan Sullivan in the analysis of data is gratefully acknowledged. This work was performed in the Laboratory for Nuclear Science and Engineering and was supported by the Bureau of Ships and the Office of Naval Research.

[24] W. S. Snyder, Nucleonics 6, 49 (Feb. 1950).
[25] Segrè Chart, revised April 1948, *The Science and Engineering of Nuclear Power*, Vol. II (1948).
[26] J. O. Hirschfelder and E. N. Adams, Phys. Rev. 73, 863 (1948).
[27] P. C. Gugelot and M. G. White, Phys. Rev. 74, 1215 (1948).

Storage of Small Signals on a Dielectric Surface

JOHN V. HARRINGTON
Air Force Cambridge Research Laboratories, Cambridge, Massachusetts
(Received May 29, 1950)

A mathematical analysis is presented which is believed to be applicable to a general class of storage tubes where signal storage is accomplished by depositing through secondary emission a charge pattern on a dielectric surface. The assumptions made to linearize and simplify the problem are outlined and plots are given of the predicted output signals for writing, reading, and cancellation operations when the input signal is a step function. Experimental evidence is presented to substantiate the analytical results.

INTRODUCTION

AN increasing amount of material has appeared in recent literature concerning the general class of vacuum tubes called storage tubes, in which signal storage is accomplished by depositing, through secondary emission, a charge pattern on a dielectric surface.[1–5] The material published on these storage tubes thus far has emphasized more or less the physics of the storage process—an emphasis which is to be expected in view of the admittedly complicated process of charge transfer by secondary emission and the difficulty in collecting the secondary current to give a useful output signal. To the engineer using these storage tubes, as opposed to the tube designer, somewhat different characteristics are of primary interest. As in ordinary vacuum tubes, these are external or equivalent characteristics and are defined primarily on the basis of the device's behavior as a circuit element; i.e., on the basis of its effect on the amplitude or shape of a signal passed through it.

The analysis which follows attempts to explain quantitatively some of the external characteristics of the *barrier-grid storage tube*, and a substantial portion of the introductory discussion concerns this particular model. The results which are obtained, however, are believed to be sufficiently general as to apply to the general class of tubes under discussion.

A simplified sketch of the *barrier-grid tube* is shown in Fig. 1. A complete description of this device and its operation appears in the literature.[1] In operation the dielectric surface will be charged, in the absence of a signal, to some potential V_0 at which the effective secondary emission ratio, δ, equals 1 and equilibrium exists. Thereafter when a signal is applied to the back plate, the emission ratio is changed; this results in a charge, different from the equilibrium charge, being deposited on the dielectric at the point of beam impingement. Thus, for a linear beam motion a signal varying in time would be stored as a charge distribution varying in distance along a line of the dielectric surface. In reading, when no signal is applied to the back plate, the beam in sweeping over this same line acts to restore the charge distribution to the equilibrium value, and in so doing the secondary emission current (as observed at the collector) will regenerate essentially a counterpart of the stored signal.

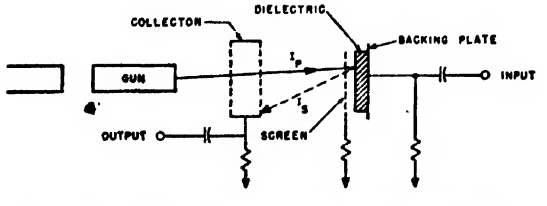


FIG. 1. Barrier grid storage tube, simplified sketch.

BASIS OF ANALYSIS

A rigorous mathematical analysis of a device such as this is extremely difficult if not impossible to make. Consequently certain assumptions and simplifications must be made to bring the problem within the range of available analytical methods. In this case the following assumptions have been made:

(1) The loss of resolution due to the redistribution of secondaries is neglected. The screen action is assumed to reduce such losses to negligible proportions.

(2) Not all of the secondary beam leaving the dielectric gets to the collector. It is assumed, however, that the fraction of the secondaries which are intercepted by the screen and other electrodes is constant; thus the collector or signal current is a "scaled-down" version of the total secondary current.

(3) A linear relationship is assumed between small voltage changes about the second crossover, or equilibrium, point of the secondary emission characteristic, and the corresponding secondary emission ratio (Fig. 2).

(4) The current in the electron beam is assumed to follow an even distribution curve in the direction of motion of the beam and

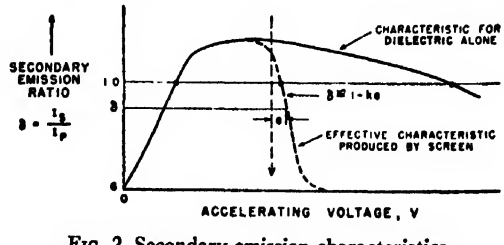


FIG. 2. Secondary emission characteristics, with and without screen.

[1] Jensen, Smith, Mesner, and Flory, RCA Review (March, 1948).
[2] R. A. McConnell, Proc. I.R.E. (November, 1947).
[3] A. V. Haeff, Electronics (September, 1947).
[4] H. Klemperer, J. T. de Bettencourt, Electronics (August, 1948).
[5] V. K. Zworykin, G. A. Morton, Television (John Wiley and Sons, Inc., New York, 1939).

to be uniformly distributed in the transverse direction. The effect of this assumption is to reduce what would be a three-dimensional problem to one of two dimensions—distance and time. This may be done with small sacrifice of generality, since it is primarily the lateral and not the transverse distribution of the beam current which influences the signal response. For purposes of presenting the results of this analysis graphically a normal beam distribution is assumed.

(5) Effects such as unevenness or non-uniformity of the dielectric, secondary emission from the screen material, etc., which contribute primarily to the background noise level of the tube, are ignored.

The foregoing assumptions allow us to reduce the problem to that of evaluating the current-voltage relationships in the equivalent network shown in Fig. 3. $e_0(t)$ is the signal voltage on the back plate applied at time $t=0$; C_x is the capacitance of the dielectric per unit area; and δ is the effective secondary emission ratio and is given by $\delta=1-ke$ for small values of e.

The quantity I_x is the primary beam-current density at point x; if we assume that the beam is scanning with constant lateral velocity v, then I_x as a function of x

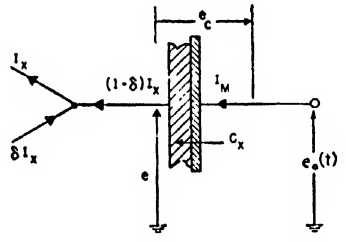


FIG. 3. Storage-tube input circuit.

and t is given by:

$$I_x=\frac{I_b}{w}F\left(\frac{x-x_0-vt}{w}\right)=\frac{I_b}{w}F(x'-t')$$

where $F(x')$ is an even function of x' and meets the conditions:

$$\int_{-\infty}^{\infty}F(x')dx'=1 \text{ so that } \int_{\infty}^{\infty}Idx=I_b$$

and for $|x'|\gg 1$, $F(x')\to 0$.

I_b is the total beam current per unit line width and w is a measure of a half beam width. The primed variables are normalized or dimensionless units of distance and time are given by

$$x'=(x-x_0)/w, \quad t'=vt/w.$$

For the circuit of Fig. 4 we may write the following equations

$$e=e_0(t)-e_c, \tag{1}$$

$$C_x(de_c/dt)=(1-\delta)I_x. \tag{2}$$

Then the basic differential equation obtained by combining the foregoing relationships is

$$de_c/dt'+\gamma F(x'-t')e_c=\gamma F(x'-t')e_0(t)$$

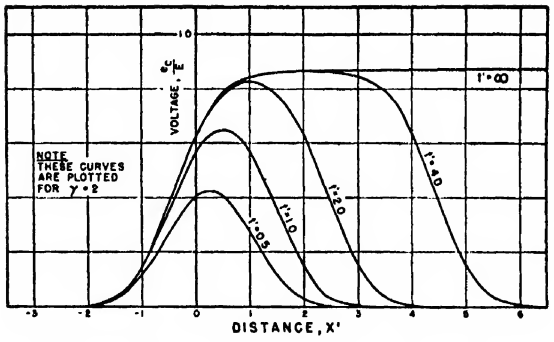


FIG. 4. Build-up of voltage distribution along scan line during storage of step voltage.

where

$$\gamma=[kI_b/vC_x]. \tag{3}$$

Since the signals to be stored in any practical case are most always pulsed signals, it is useful to carry out the analysis on the assumption that $e_0(t)$ is a step function of magnitude E applied at $t=0$. From these results one can easily deduce the tube performance for pulsed signals. Under these conditions (3) has the general solution

$$e_c(x', t')=E+C_1(x')\epsilon^{\gamma[B(x'-t')-B(x')]}, \tag{4}$$

where

$$B(x')=\int_0^{x'}F(x')dx'$$

and $C_1(x')$ is a constant of integration (the integration having been performed with respect to t') to be determined by the initial conditions.

The varying component of the collector current which generates the output signal will be the negative of the dielectric charging current I_M; it is calculated as follows:

$$I_M=\int_{-\infty}^{\infty}(1-\delta)I_xdx=\int_{-\infty}^{\infty}C_x\frac{de_c}{dt}dx=vC_x\int_{-\infty}^{\infty}\frac{de_c}{dt}dx. \tag{5}$$

In setting the limits for which this integration is to be performed, the implication is made that the length of a scan line is infinite. This is a valid assumption so long as the scan line is very long compared to one beam width.

ANALYSIS

It is both convenient and useful to obtain solutions for the charging current and for the voltage distribution along a scan line for three general cases. The first case will correspond to the so-called writing operation in which a voltage step is to be stored with no signal initially in storage; the second case will correspond to a multiple writing or cancellation operation in which a second-step voltage is to be stored over one initially in storage; the third case corresponds to a reading operation in which no external signal is applied but a stored signal is present.

Case I. Writing Operation

For this case the initial conditions are at $t=0$, $e_c=0$, $e_0(t)=E$, when these are used to evaluate $C_1(x)$, the desired expression for e_c is:

$$e_c(x', t')=E[1-\epsilon^{-\gamma[B(x')-B(x'-t')]}]. \quad (6)$$

It is interesting to plot this (Fig. 4) for an assumed[6] normal beam current distribution in order to observe the changing charge pattern laid down by the electron beam as the applied step voltage is stored. In interpreting these curves it is helpful to consider t' as a distance giving the center of the beam. For very large values of t', the voltage distribution especially in the vicinity of $x'=0$ becomes essentially independent of time and is given by

$$\operatorname*{Lim}_{t'\to\infty} e_{c1}(x', t')=e_{c1}(x', \infty)=E_1[1-\epsilon^{-\gamma[\frac{1}{2}+B(x')]}]. \quad (7)$$

This expression represents the "stored" signal.

Case II. Multiple Writing Operation

By using (7) as the initial voltage distribution for the second writing operation in which a step of magnitude E_2 is to be stored, it is found that

$$e_{c2}(x', t')=E_1[1-\epsilon^{-\gamma[\frac{1}{2}+B(x')]}]\epsilon^{-\gamma[\frac{1}{2}-B(x'-t')]}; \quad t\leq 0 \quad (8)$$

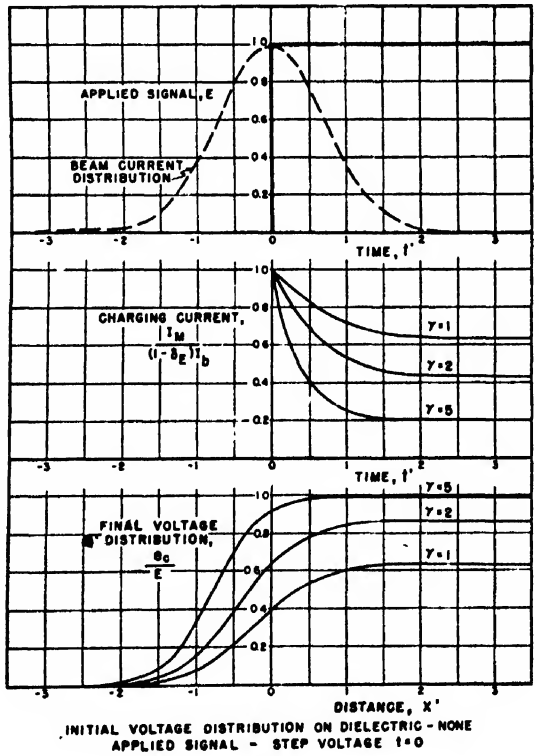


FIG. 5. Signal storage on a dielectric surface, Case I.

[6] That is, in the curves, $F(x')=[I_b/(\pi)^{\frac{1}{2}}w]e^{-x'^2}$, and $B(x')=\frac{1}{2}\operatorname{erf}(x')$.

$$e_{c2}(x', t')=E_2[1-\epsilon^{-\gamma[B(x'-t')-B(x')]}]+E_1\epsilon^{-\gamma[\epsilon^{\gamma[\frac{1}{2}+B(x')]}-1]}\times\epsilon^{\gamma[B(x'-t')-B(x')]}; \quad t\geq 0. \quad (9)$$

The final "stored" signal obtained by letting $t'\to\infty$, is then

$$e_{c2}(x', \infty)=(E_2+\epsilon^{-\gamma}E_1)[1-\epsilon^{-\gamma(\frac{1}{2}+B(x'))}]. \quad (10)$$

From (8), (9) and (5) one finds that the charging current is given by

$$I_M=kE_1I_b\cdot\epsilon^{-\gamma}\left[I(\gamma, t')-\left(\frac{\epsilon^{\gamma}-1}{\gamma}\right)\right]; \quad t\leq 0 \quad (11)$$

$$I_M=kI_b\cdot\left[(E_2+E_1\epsilon^{-\gamma})I(\gamma, t')-E_1\left(\frac{1-\epsilon^{-\gamma}}{\gamma}\right)\right]; \quad t\geq 0 \quad (12)$$

where the function $I(\gamma, t')$ is given by:

$$I(\gamma, t')=\int_{-\infty}^{\infty}\epsilon^{\gamma[B(x')-B(x'-t')]}F(x')dx'. \quad (13)$$

Equations (11) and (12) are rather general expressions for the dielectric charging current and as special cases of these one may obtain by setting $E_1=0$, $E_2=E_1=E$, or $E_2=0$, the expression for I_m for cases I, II, or III,

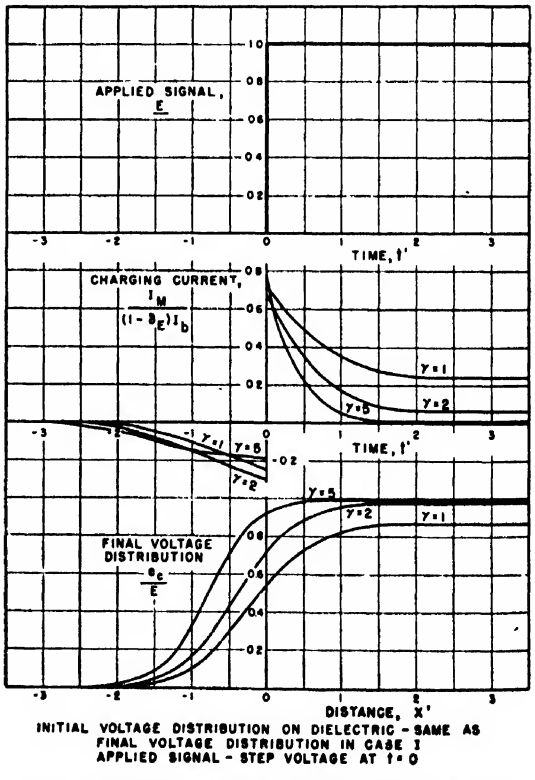


FIG. 6. Signal storage on a dielectric surface, Case II.

respectively. It is interesting to note that while a simple relation giving $I(\gamma, t)$ is in general impossible to obtain because of the difficulty encountered in integrating (7), its values at $t=0$, and $t=\pm\infty$ can be calculated.

$$I(\gamma, 0)=1$$
$$I(\gamma, \infty)=(1-\epsilon^{-\gamma})/\gamma$$
$$I(\gamma, -\infty)=(\epsilon^{\gamma}-1)/\gamma. \quad (14)$$

Thus we may conclude that the particular beam distribution, i.e., the particular $F(x')$; only influences the shape of the response within a few beam widths travel from the point of application of the signal.

Even within this region the $I(\gamma, t')$ functions for different assumed beam current distributions are very much alike. Beyond this region, the response (a sort of steady state response) is independent of $F(x')$ and depends principally on γ. This is to be expected since for a flat or uniform response only the total beam current and not its distribution should be significant.

The charging current and final voltage distribution for selected values of γ and an assumed normal beam current distribution for cases I and II are plotted[7] in Figs. 5 and 6. The typical "spike" for the writing signal is indicated, and it is seen that the effect of the finite width of the electron beam is to slope the leading edge of the stored signal such that two modified distance units (roughly, one beam width) are required to go from initial to final values.

Case III. Reading Operation

The expressions for charge distribution and charging current for the reading operation may be easily obtained from (10)–(12), by setting $E_1=E$ and $E_2=0$. A comparison of (10) and (7) indicates that for the reading operation the initial and final voltage distributions would be identical except for a scale factor $\epsilon^{-\gamma}$. Thus the reading operation is also an erasing operation in which the stored signal is reduced by $\epsilon^{-\gamma}$ or $8.68\gamma db$.

Typical curves for the Case III charging current are plotted in Fig. 7 along with the corresponding voltage distribution curves for the reading operation. It is seen that the output current wave rise time has been limited to about 3-to-4 modified time units or to the time it takes the beam to sweep through 1.5 to 2 beam widths.

EXPERIMENTAL RESULTS

As intimated previously the purpose of this section is to present some experimental evidence obtained with the STE barrier-grid storage tube which will support the preceding analysis. To obtain this evidence it is first necessary to choose a write-read schedule which on a continuous or repetitive basis will yield signal shapes corresponding to those predicted for Cases I, II and III.

[7] For selected values of γ and for an assumed normal beam current distribution, numerical values of $I(\gamma, t')$ in the range $-4<t'<4$ have been calculated at the M.I.T. Center of Analysis, Cambridge, Massachusetts.

It turns out that about the simplest schedule to do this is two writing followed by two reading operations, ad infinitum (Fig. 8). Because of the repeated application of the signal and the improbability of complete erasure, the initial charge distribution prior to the application of the first writing pulse will not be zero as assumed in Case I but will have some finite value $AE(x)$. It is desirable to calculate the magnitude of A as a function of γ so that our results may be properly interpreted. Thus, from the preceding analysis,

if before *first write*, stored signal is $A\cdot E(x')$ (15)

then before *second write*, stored signal is
$$(1+\epsilon^{-\gamma}A)\cdot E(x') \quad (16)$$

and before *first read*, stored signal is
$$[1+\epsilon^{-\gamma}(1+\epsilon^{-\gamma}A)]\cdot E(x') \quad (17)$$

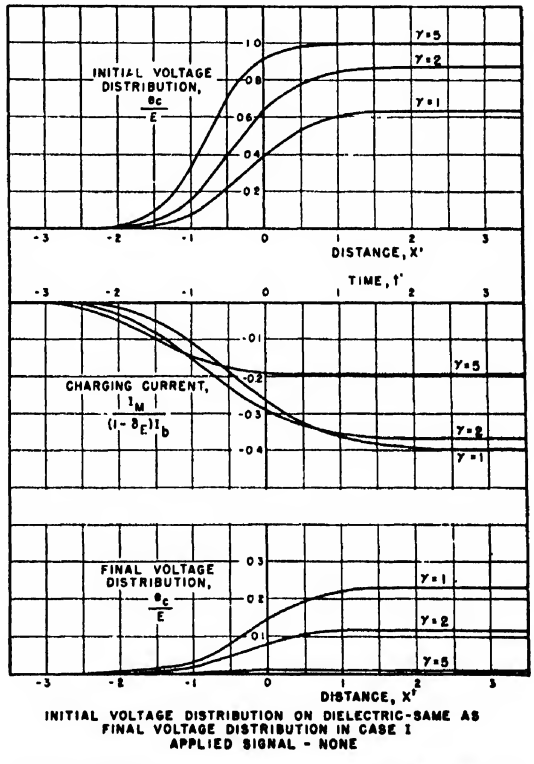


FIG. 7. Signal storage on a dielectric surface, Case III.

and before *second read*, stored signal is
$$\epsilon^{-\gamma}[1+\epsilon^{-\gamma}(1+\epsilon^{-\gamma}A)]\cdot E(x') \quad (18)$$

and after *second read*, stored signal is
$$\epsilon^{-2\gamma}[1+\epsilon^{-\gamma}(1+\epsilon^{-\gamma}A)]\cdot E(x'). \quad (19)$$

Then for a sustained sequence of such operations equilibrium will be reached when (15)=(19), or when

$$A=\epsilon^{-2\gamma}\frac{(1+\epsilon^{-\gamma})}{(1-\epsilon^{-4\gamma})}. \quad (20)$$

Using this to determine the initial values of the charge distribution, one can utilize the results of the previous

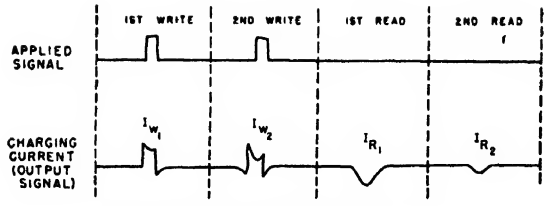


FIG 8 Two write two read schedule.

section, Eq. (12), to find the steady-state responses of the write-read signals of Fig. 8. When this is done one finds that

$$I_{w1}(\infty)=I_{R1}(\infty)=(1-\delta_E)I_b\cdot\frac{(1-\epsilon^{-\gamma})}{\gamma(1+\epsilon^{-2\gamma})} \qquad (21)$$

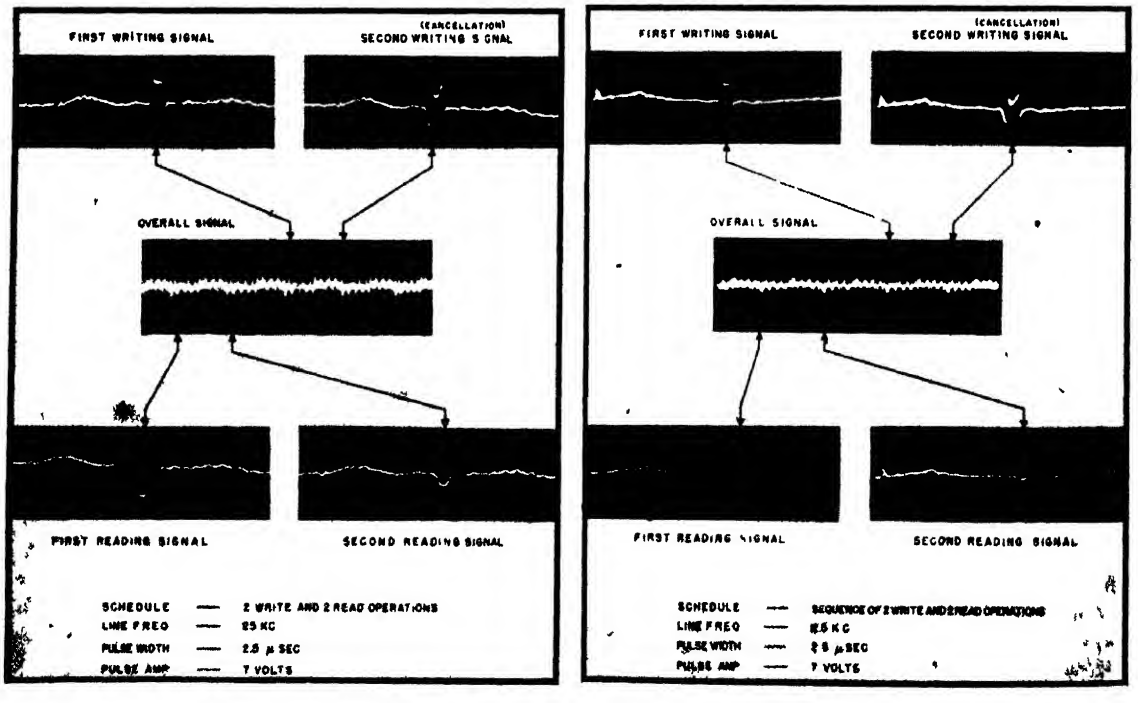
and

$$I_{w2}(\infty)=I_{R2}(\infty)=(1-\delta_E)I_b\,\frac{\epsilon^{-\gamma}(1-\epsilon^{-\gamma})}{\gamma(1+\epsilon^{-2\gamma})}. \qquad (22)$$

For a value of γ in the neighborhood of unity, the magnitude of A will be small, and hence the initial conditions for the first writing operation will be essentially the same as those assumed in Case I. Similarly, the basis for comparison of the second writing signal and the reading signal with Cases II and III, respectively, is also established.

Oscillograms of the writing and reading signals for a schedule as outlined above are shown (Figs. 9(a) and (b)) for $\gamma_{avg}=0.8$ and $\gamma_{avg}=1.6$. In general, a very good agreement is noted between these signals and their predicted counterparts in Figs 5–7 It was necessary here to define the term γ_{avg} as the average of the writing and reading decrements, since for a 7-volt signal the effective secondary emission characteristic is not too linear or even symmetrical. This may be seen from Fig. 10, in which the initial amplitude of a writing signal (followed by three reading operations to insure essentially complete erasure for $\gamma\cong 1$) is taken as a measure of $(1-\delta)I_b$ and is plotted, in arbitrary units, *versus* the input voltage. The assumption of linearity is evidently valid over a range of about ± 5 volts. This is further substantiated by a plot of the *first* reading signal against the input pulse amplitude, which gives a sort of input-output characteristic for the tube. Here again it is seen that for small input signals ($<\pm 5$ volts) one may consider the operation to be approximately linear.

While these oscillograms give a fair measure of confirmation to our analysis, it is desirable to obtain more quantitative support for this theory. This may be obtained by comparing observed steady-state signal amplitudes with the theoretical amplitudes given by (21) and (22). (See Fig. 11.) It will be noted that the experimentally observed amplitudes agree fairly well with the calculated amplitudes. It should be mentioned that the sets of points were obtained by some rather inaccurate measurements of write and read signal amplitudes on a synchroscope as the average γ was changed by varying the velocity of the storage-tube sweep or scan. Furthermore, in order to provide a basis of comparison the first reading in each set was scaled to agree with the theoretical value for that particular γ and the same scale factor applied to the remaining



(a) $\gamma_{avg}=0.8$ (b) $\gamma_{avg}=1.6$

FIG. 9. Storage-tube output signals.

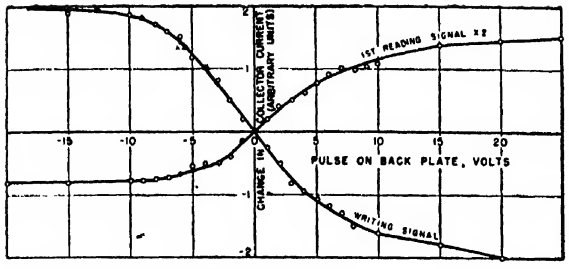


FIG. 10. Amplitude of writing signal and first reading signal *vs.* input pulse amplitude.

observations in the set. Thus, no comparison of absolute values is made here, but rather a comparison of the patterns or curves along which these points fall; from these results one then draws the conclusion that a rather good correlation between predicted and observed curves exists. It might be mentioned again that these amplitudes are difficult to measure with any great degree of accuracy, especially when the applied signal is kept low (at about three-volt peak) to insure good linearity since the base-line disturbance or background "noise" generated by the tube thus appears proportionately larger. Hence, the data of Fig. 11 indicate that the analysis is correct to at least a good first approximation.

Further support for the analysis is obtained by plotting experimentally observed values[9] of γ *versus* the reciprocal of the scanning velocity for various values of collector current (Fig. 12). It will be noted that the linear relationship between γ and $1/v$ is confirmed and also to a more approximate degree the direct variation of γ with I_b.

SUMMARY AND CONCLUSIONS

It will be noted that the shape of the stored voltage distribution in each of the three cases considered is the same, differing only in magnitude, the effectiveness of the erasing action in Case III being measured by the factor $\epsilon^{-\gamma}$ and the scale factor in Case II being the sum of those in I and III. This is a very useful result, since it means that for sufficiently high values of γ (say, above 2) the final value of the stored voltage distribution is dependent only upon the applied signal and is practically independent of the initial charge distribution on the dielectric. Therefore, under these conditions the signal being stored takes precedence over any previously in storage and no separate erasing operation is required.

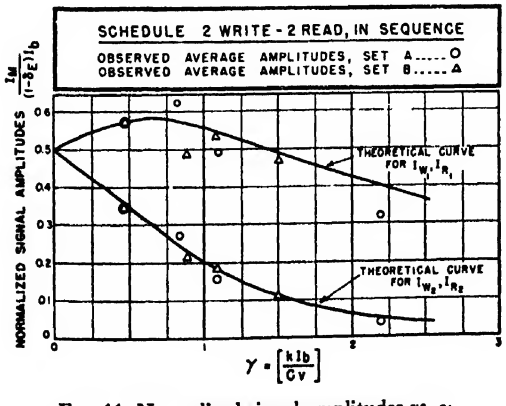


FIG. 11. Normalized signal amplitudes *vs.* γ.

[9] This data was obtained by T. F. Rogers and J. Dumanian.

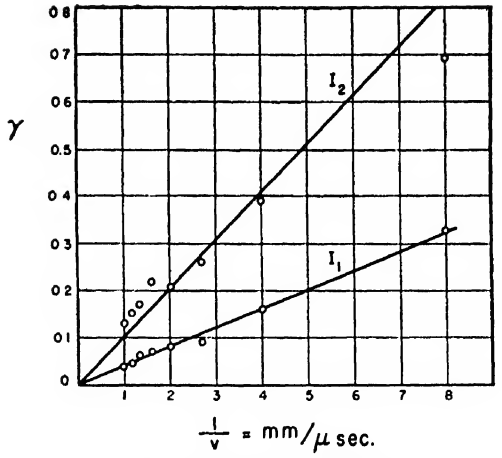


FIG. 12. γ *vs.* $(1/v)$ (for two values of collector current).

The factor $\gamma = kI_b/vC_x$ is evidently an extremely important quantity in governing the tube operation; the reason for this may be surmised from an understanding of its physical meaning. The expression for γ may be written as

$$\gamma = \frac{kI_b}{vC_x} = \frac{(1/v)}{(C_x/kI_b)} = \frac{(1/v)}{R_eC_x},$$

where $R_e = 1/kI_b$. This interpretation is dimensionally valid since k, the slope of the secondary emission curve, has the dimensions of $(\text{volts})^{-1}$; therefore $1/kI_b$ will have the dimensions of ohms multiplied by length. R_eC_x may be interpreted to be the time constant for charging the dielectric in the absence of beam motion and hence, γ is seen to be the ratio of the time it takes the beam to traverse a given spot to the time it would take the beam to charge the same spot if the beam were stationary. For reasonably successful charging action it would be expected that these times should be somewhere near equal, or that the important range of γ-values should be in the vicinity of unity; this expectation is borne out by the foregoing analysis.

ACKNOWLEDGMENTS

The author wishes to express his appreciation for the help and criticisms received in this study from many of the members of the Air Force Cambridge Research Laboratories and from Dr. A. S. Jensen of the RCA Laboratories, Princeton, New Jersey.

Graphical Representation of Particle Trajectories in a Moving Reference System

M. GARBUNY
Westinghouse Research Laboratories, East Pittsburgh, Pennsylvania
(Received April 17, 1950)

A graphical method is derived for the analysis of microwave electron tubes, ion accelerators etc., which refers particle positions and velocities to a moving reference system. If the forces are dependent on time only, the trajectories are transformed into straight lines. For inhomogeneous fields an approximation procedure applies. To demonstrate the capabilities of this method a brief treatment of the transit time phenomena in cavity triodes is outlined.

I. INTRODUCTION

THE analysis of transit time phenomena of particles under the influence of varying forces, as in electron tubes, leads often to complicated transcendental equations. Usually cumbersome graphical methods are employed to arrive at such kinematic values as arrival velocities, transit angles, and current distribution.[1] A typical example is the simple case of linear electron motion between two electrodes which produce superimposed homogeneous static and radiofrequency fields, i.e. a type of acceleration $a(t)$ described by

$$a(t)=A\sin t+B, \tag{1}$$

t being measured in electrical radians. An electron leaving the first electrode at time t_1 with a velocity v_0 arrives at a distance s at time t with a velocity v. One obtains by successive integrations:

$$v=A(\cos t_1-\cos t)+B(t-t_1)+v_0, \tag{2}$$

$$s=A[(t-t_1)(\cos t_1+v_0/A)+\sin t_1-\sin t]+B(t-t_1)^2/2. \tag{3}$$

The usual problem is then to evaluate transit times $(t-t_1)$ and velocities for a given electrode distance s with various starting times t_1 as parameter, and it is evident that in such representation elaborate graphical methods are called for such as a point-by-point construction of all the trajectories $s(t)$ with various values t_1.

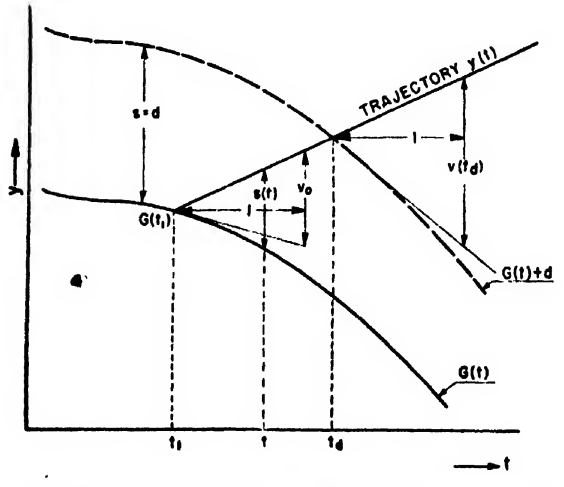


FIG. 1. Method of obtaining straight lines as trajectories.

In the following a graphical method will be described which in essence amounts to referring all trajectories to a coordinate system which moves along with one of the particles which is chosen arbitrarily for greatest simplicity of the procedure. If the force is a function only of time, as in case of homogeneous fields, all trajectories become straight lines. This means in terms of expediency that instead of constructing a great number of complicated trajectories or performing equivalent graphical operations one has to use only one relatively simple trajectory as a reference system, and all the others appear as straight lines, permitting a simple evaluation of such values as were mentioned above.

II. DERIVATION OF METHOD

Let us assume a particle enters a field at time t_1 with a velocity v_0, and the forces acting on it depend only on time so that the resulting acceleration is

$$a=f(t). \tag{4}$$

Let the motion be one-dimensional for reasons of simplicity, although the three-dimensional case can be obtained by considering the components $f_x(t)$, $f_y(t)$, $f_z(t)$ separately. We now define:

$$F(t)=-\int_p^t f(t)dt;\quad G(t)=\int_q^t F(t)dt, \tag{5}$$

where the lower limits p and q of the integrals can be chosen arbitrarily or preferably so that the simplest expressions for F and G result. One obtains then for velocity and distance reached at time t:

$$v=F(t_1)-F(t)+v_0=v_0+G'(t_1)-G'(t), \tag{6}$$

$$s=G(t_1)-G(t)+(v_0+G'(t_1))\cdot(t-t_1). \tag{7}$$

If one writes (7) in the following form as two equations

$$G(t)+s(t)\equiv y(t)=(v_0+G'(t_1))(t-t_1)+G(t_1) \tag{8}$$

one obtains two statements. The left side of (8) defines $y(t)$ as a transformed trajectory of position *vs.* time by the stipulation that

$$s(t)=y-G, \tag{9}$$

[1] A notable exception is R. Kompfner's method, Wireless Engr. 19, 2, (1942), which for a special case overlaps with the method described here.

i.e. that the distance traveled equals the vertical distance between the trajectory and the reference curve $G(t)$. The right side of (8) states that the trajectory $y(t)$ is a straight line of slope $m=v_0+G'(t_1)$ through the point with the coordinates t_1, $G(t_1)$. The velocity at each point

$$v(t)=ds/dt=y'(t)-G'(t) \qquad (10)$$

equals the difference between the slopes of the trajectory $y(t)$ and the reference curve $G(t)$. The physical interpretation of this procedure follows now easily. $G(t)$ is the trajectory of a selected particle in rectangular coordinates. With (9) and (10), one in effect refers positions and velocities to that of the selected moving particle, thereby attaining straight lines for the newly defined trajectories $y(t)$.

One thus arrives at a simple graphical method for the analysis of transit time phenomena, see Fig. 1. First the reference curve $G(t)$ is plotted, as defined by (4) and (5). Next, the straight line trajectories are drawn starting from $G(t)$ at various points t_1, $G(t_1)$. The slope of the trajectory equals the sum of the slope of the curve $G(t)$ at point t_1 and the initial velocity v_0 at this point; if the particle starts at rest, a tangent results. The intersection of the trajectory with a curve parallel to $G(t)$ at a distance $s=d$ marks the arrival time t_a. The arrival velocity is given by the vertical distance of the trajectory from the tangent to the curve at the intersection, at distance one (unit of the abscissa) from this intersection. Other kinetic values can be obtained by simple additional constructions.

A review of the literature on pertinent graphical methods revealed that for the special case $f(t)=A\sin t$ the described transformation is identical with Kompfner's procedure,[1] which however arrives at parabolic trajectories for the more general case of (1), $f(t)=A\sin t+B$. Another special case is $f(t)=0$, $G(t)=0$ for field-free drift spaces which leads to Applegate's diagram.[2]

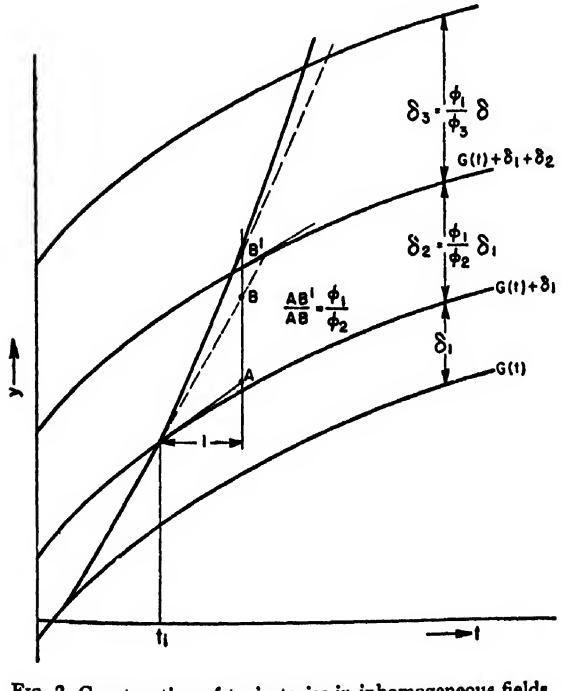


FIG. 2. Construction of trajectories in inhomogeneous fields.

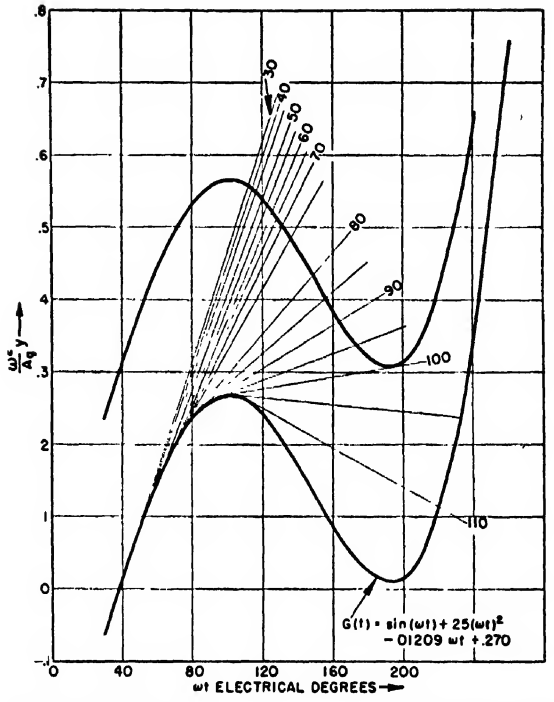


FIG. 3. Transformed electron trajectories for idealized conditions of cathode-grid cavities in dimensionless units. The chart results from the choice of $B_0/A_0=0.50$, $\delta_0=0.298$ (e.g., $d/\lambda=3.85\times10^{-3}$, $V_0=1$ Kv) and a p and q in (5) so as to increase the accuracy of the graph. The value for δ_0 corresponds to the limitation that the first emitted electron and the last not reflected electron leave at 30 and 100 electrical degrees, respectively.

III. EXTENSION OF THE METHOD TO INHOMOGENEOUS FIELDS

The method described above postulated forces which depended on time only. In inhomogeneous fields however the particle experiences forces which, in addition to being time variant, are also functions of position, i.e. one obtains for the acceleration instead of (4) the more general differential equation

$$d^2x/dt^2=\varphi(x)f(t). \qquad (11)$$

Here $\varphi(x)$ is, for one-dimensional motion, the static field distribution along the path of the particle. In simple geometries $\varphi(x)$ can be computed, while in other cases it has to be determined empirically, e.g. with the electrolytic tank method. One adopts now the following approximation method for the transformation of the particle trajectories, see Fig. 2.

The traversed distance $x=d$ is subdivided into a number n of intervals Δx_λ which are equal and small enough so that within them the variation of $\varphi(x)$ from

[1] See, e.g., *Klystron*, Sperry Gyroscope Company, 1944.

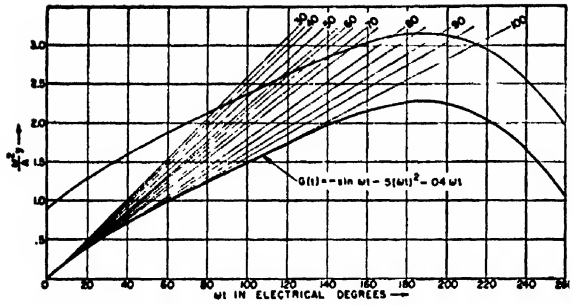


FIG. 4. Trajectories in the anode-grid cavity of a triode. Phase values marking the individual electron traces refer to the times of starting at the cathode in Fig. 3. Phase notation of the abscissa refers to the r-f voltage in the anode gap which adjusts itself to the bunched electron stream. $B/A=1$, $\delta=0.831$ (e.g., $d/\lambda=0.02$, $V=10$ Kv). A conversion efficiency of 62 percent is found by an evaluation of arrival velocities.

an average value φ_λ is sufficiently small. The condition that the forces are only time dependent is therefore approximately fulfilled within each segment.[3] The reference curves $G(t)$ and $G(t)+\delta_1$, are then plotted for the first interval Δx_1, $G(t)+\delta_1+\delta_2$ as the upper limit of the second interval Δx_2 etc. so that one plots altogether $n+1$ parallel traces of the same curve $G(t)$. The distances between the curves are inversely proportional to the fields in the intervals which they confine, i.e. $\delta_{\lambda+1}/\delta_\lambda=\varphi_\lambda/\varphi_{\lambda+1}$. The trajectories are straight lines within each interval, but at the crossing of a reference line the slope is changed. It can easily be verified that the slopes of two adjoining trajectory segments, when referred to the slope of the reference line at the intersection, are again inversely proportional to the field regions they traverse, i.e., referring to Fig. 2, $(y_{\lambda+1}'-G'(t_i))/(y_\lambda'-G'(t_i))=AB'/AB=\varphi_\lambda/\varphi_{\lambda+1}$, where $G'(t_i)$ is the slope of the reference curve at the intersection point. One obtains in this manner an approximation to the transformed trajectories in inhomogeneous fields.

[3] If the maximum variation of $\varphi(x)$ from φ_λ within Δx_λ is $d\varphi_\lambda$, the difference in the results obtained by using φ_λ and $\varphi_\lambda+d\varphi_\lambda$ is certainly larger than the actually occurring error. The difference of arrival times t_λ at the end of the interval is derived from (6) and (7) as $dt_\lambda=-d\varphi_\lambda(\Delta x_\lambda-v_{\lambda-1}(t_\lambda-t_{\lambda-1}))/v_\lambda\varphi_\lambda$. Hence the relative difference in transit time $dt_\lambda/(t_\lambda-t_{\lambda-1})\cong-bd\varphi_\lambda/\varphi_\lambda$, where $b=[\Delta x_\lambda-v_{\lambda-1}(t_\lambda-t_{\lambda-1})]/v_\lambda(t_\lambda-t_{\lambda-1})$ can be determined from the diagram. The relative difference in arrival velocity is then $dv_\lambda/v_\lambda-v_{\lambda-1})\cong d\varphi_\lambda(1-b)/\varphi_\lambda$. The actual errors will lie between zero and these upper limits.

IV. EXAMPLE: ELECTRON TRAJECTORIES IN CAVITY TYPE TRIODES

As a typical problem to which this method can be applied, the analysis of transit time phenomena in microwave triodes is briefly outlined in the following. The acceleration under negative bias conditions in the cathode-grid space is given by $f(t)=A_g\sin(\omega t)-B_g$. Using (5) and dividing by A_g/ω^2 to arrive at a dimensionless scale, one plots $G(t)$ (see Fig. 3) and a second parallel curve at a distance $\delta_g=\omega^2 d/A_g=2(2\pi d/\lambda\beta_g)^2$ wherein d/λ is the grid-cathode distance in fractional wave-lengths and β_g the fractional lightspeed of an electron corresponding to the peak r-f voltage $V_g=(m/e)A_g d$ across the gap. Since one usually chooses $B_g/A_g=0.5$ one such chart with varying parameter δ_g can describe most grid conditions which occur in the practical application of microwave tubes.

The velocities and times at which the electrons arrive at the grid are then applied as the initial conditions for the trajectories in the grid-anode space (Fig. 4). The square of the arrival velocities at the anode is proportional to the power loss, and one can in this manner, by averaging over all electrons, arrive at the total conversion efficiency.

The author wishes to express his appreciation for the many stimulating discussions with Dr. J. W. Coltman and Dr. W. Altar.

Some Observations on the Epitaxy of Sodium Chloride on Silver*

G. W. Johnson
Brookhaven National Laboratory, Upton, Long Island, New York
(Received May 22, 1950)

An experimental study was made of the orienting effects of silver on sodium chloride crystallized from solution. Pronounced epitaxy was observed in the following cases: (1) On oriented films of silver with (100) and also with (111) planes parallel to the surface; and (2) on polycrystalline high purity silver sheet.

INTRODUCTION

THE phenomenon of epitaxy, or oriented overgrowth of one crystal on another, has been known for over a century. The interpretation of the effect in general had to await the development of x-ray and electron diffraction techniques. Perhaps the most complete systematic study, after the x-ray lattice constants of the crystals were known, was that of Royer.[1] As a result of his work, he concluded that at least two of the following three conditions must be met before oriented growth** can occur:

(1) If crystal A is oriented by crystal B, then A must have a two-dimensional lattice which is commensurate within limits with the exposed two-dimensional lattice of B. The lattice spacings of atoms of A and B in the mutual plane of contact do not need to be the same; the lattice spacing of one crystal may be a small whole-number multiple of that of the other.

(2) The ions of A must take positions which corresponding ions of B of the same sign would have occupied had the lattice of B been extended.

(3) A and B must have the same type of binding, e.g. both ionic crystals or both metals.

Subsequently, notable exceptions were found to these conclusions. Beukers[2] discovered that galena, a semiconductor, oriented sodium chloride, an ionic crystal. Sloat and Menzies[3] looked for oriented growth of ionic salts on silver but obtained epitaxy only with NH_4Br and CsCl. It is now well established that some metals condensed in a vacuum on cleaved ionic crystals under proper conditions will assume preferred orientations.[4–6] Among these is the growth of silver parallel to the structure of rock salt. Since the cleavage surface of rock salt is the (100) plane, the condensation of silver yields a film with the (100) orientation. Films prepared in this way have a twinned structure but after heating to 500°C for a few minutes, they are converted into relatively perfect single crystals having the same orientation as the substrate.[7]

The study of the epitaxy of ionic crystals deposited from solution on metals has not been given much attention. Brück,[5] and Goche and Wilman,[7] after removing the silver film from rock salt by solution of the latter in water, observed interference points in the diffraction patterns which were attributed to the presence of sodium chloride on the surface, which had not been removed by their rinsing procedures. The indicated orientation was that the (100) plane of sodium chloride was parallel to that of silver, but that the [100] direction of the salt was along the [110] direction of silver. The films apparently were not examined under a microscope. In a recent general discussion of crystal growth by G. P. Thomson[8] and an excellent survey of the literature by van der Merwe,[9] no examples were given of epitaxy between ionic crystals and metals other than those referred to here.

This paper is a description of some observations of marked epitaxy of sodium chloride on oriented films and on polycrystalline sheets of silver. It was found that orientation resulted even when the difference in atomic spacings for the two planes in contact was large. It was concluded that the orienting influence was exerted in the embryonic stages of growth.

EXPERIMENTAL PROCEDURE

A few drops of sodium chloride solution were placed on the surface of a suitable silver specimen (the preparation of which is described below), and after evaporation of the solvent, the disposition of the sodium chloride crystals was examined with a microscope.

Two types of silver surfaces were employed: oriented thin films and polycrystalline sheet. The use of thin films provided a ready means of obtaining surfaces with known crystallographic planes exposed. It was essential to have such surfaces in order to relate the observed orientation of the deposited sodium chloride crystals with that of the silver. Polycrystalline sheet was used to find in a qualitative way how universal the orienting influences were, and to determine whether the grain boundaries exerted any specific effects.

The method of Brück and others[4–6] was applied to the

* Research carried out at Brookhaven National Laboratory, under the auspices of the AEC.
[1] L. Royer, Bull. Soc. Franc. Miner. 51, 7 (1928).
** Epitaxy and oriented overgrowth will be used synonymously.
[2] M. C. F. Beukers, Rec. Trav. Chem. 58, 435 (1939).
[3] C. A. Sloat and A. W. C. Menzies, J. Phys. Chem. 35, 2005 (1931).
[4] H. Lassen and L. Brück, Ann. Phys. (Leipzig) 22, 65 (1935).
[5] L. Brück, Ann. Phys. (Leipzig) 26, (5), 233 (1936).
[6] O. Rudiger, Ann. Phys. (Leipzig) 30 (5), 505 (1937).
[7] O. Goche and H. Wilman, Proc. Phys. Soc. 51, 625 (1939).
[8] G. P. Thomson, Proc. Phys. Soc. 61, 403 (1948).
[9] J. H. van der Merwe, Disc. Faraday Soc., "Crystal Growth," No. 5, 201 (1949).

production of films with the desired orientations. Silver was vaporized in a vacuum from a tungsten filament and condensed (1) on cleaved rocksalt, and (2) on mica. To assure that properly oriented films are formed with this technique, it is necessary that the crystals be cleaved just prior to their introduction into the vacuum chamber. In addition, both the mica and the rocksalt must be held at a temperature above 150°C while the silver is being condensed. In our work the crystals were heated to 190°C during deposition. In this way, a silver film with the (100) plane exposed was produced on rock salt, and on mica the (111). Sufficient silver was evaporated to get a film thickness of ~1000A. The films were allowed to cool in a vacuum before removing them from the tube.

To reduce the possibility that the substrate might in some way influence the observations, the rock salt was dissolved in water. The film floated on the surface of the water and was readily picked up on a polished silver plate. The epitaxy experiments were made on the surface of the film that had not been in contact with the salt.

The film on mica was ready for testing for epitaxy as soon as it was taken from the vacuum chamber. It was unnecessary to remove the film from the substrate since it was shown by Royer[10] that mica does not orient sodium chloride The preparation of the polycrystalline silver sheet involved cleaning, heat-treatment, and in some cases etching. Square pieces ($\frac{1}{2}'' \times \frac{1}{2}''$) were cut from a 20-mil thick plate. These squares were washed carefully in detergent and rinsed successively in distilled water, benzene, and alcohol. The specimens were heated in a vacuum at 700°C for one hour and furnace cooled to removed the effects of the rolling of the sheet during its fabrication. Relatively large grains (~0.5 mm in diameter) resulted from this treatment. Each specimen then was electrolytically polished using published methods.[11,12] Although polishing was not essential to the production of epitaxy, a polished surface was easier to examine. To reveal the grain boundaries some specimens were etched in a mixture of equal parts of five percent solutions of potassium cyanide and ammonium persulfate.

The sodium chloride was crystallized from saturated solutions in mixtures of water and ethyl alcohol. By varying the amount of alcohol the rate of evaporation could be changed permitting the control of crystal size over wide limits. If large crystals were desired, an aqueous solution was necessary. On the other hand, if extremely small crystals were required, a strongly alcoholic solution was used.

To test for epitaxy, a few drops of the solution were placed on the surface. The surface was then dried under an electric hair-dryer. On a grease-free surface, the crystallization was spread uniformly over the surface except for variations from grain to grain due to orientation effects.

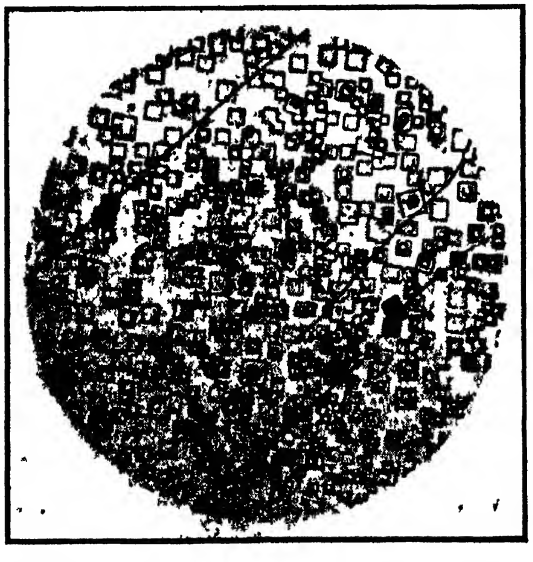


FIG 1 Epitaxy of sodium chloride on a silver film oriented by rock salt. The surface was a (100) plane The lines extending across the film were produced by cleavage steps on rock salt surface 300×

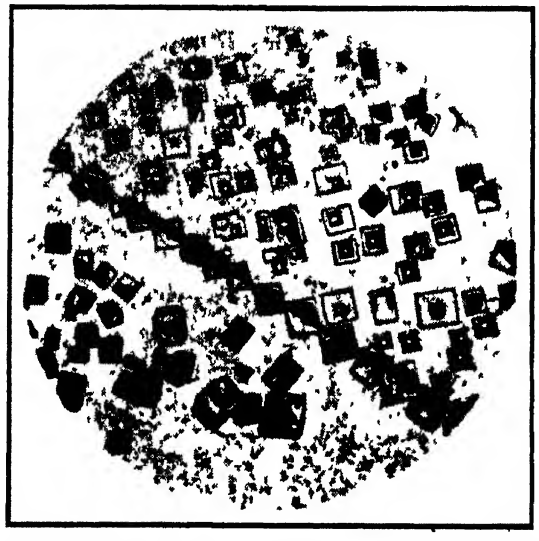


FIG 2 Sodium chloride crystals near edge of silver film oriented by rock salt Silver film supported by a polycrystalline silver plate. Note growth of crystals from edge of film 300×

Each specimen was then examined under a microscope to observe the presence or absence of preferred orientations. If the crystal size was unfavorable, the surface was cleaned by rinsing in distilled water. Another solution then was tried to get the desired crystal size. This procedure could be repeated many times, although usually after several rinses the surface would show evidence of contamination, as indicated by failure of the solution to wet the surface completely. Representative

[10] L. Royer, Comptes Rendus 194, 620 (1932).

[11] Shuttleworth, King, and Chalmers, Metal Treatment 14 (No. 51) 161 (1947).

[12] L. I Gilbertson and O. W. Fortner, Trans. Electrochem. Soc. 81, 199 (1942).

arrangements were photographed using a Bausch and Lomb metallograph under bright field vertical illumination.

Material

The silver was of high purity (99.999 percent) supplied by Sigmund Cohen. The rock-salt crystals were cleaved from a synthetic single crystal from the Harshaw Chemical Company. The mica was ordinary muscovite.

RESULTS

Thin Films—On the films with the (100) plane exposed, almost complete orientation of the sodium chloride was achieved. Characteristic results are shown

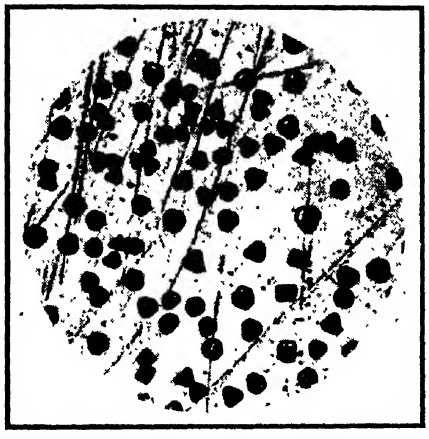


FIG. 3. Sodium chloride crystallized on silver film oriented by mica. 400×

in Fig. 1. The crystals formed rectangular pyramids whose bases were in contact with the silver surface. The straight edges forming the sides of the squares were [100] directions of sodium chloride and were directed along the [110] of the underlying silver. Thus the (100) planes of the salt and the silver were parallel, but the cube edge of the salt was aligned with the face diagonal of the silver.

Since the silver film was mounted on a polycrystalline silver plate, it was of interest to compare the areas of the film and the plate adjacent to the edge of the film to determine if the plate was affecting the results. Such an edge is shown in Fig. 2. On top of the film proper the sodium chloride crystals were oriented as illustrated in Fig. 1, whereas on the plate no general preferred alignment was noted. The edge of the film was the [100] direction of silver. The rotation of the salt crystals by 45° with respect to cube edge of the silver is plainly indicated. Of additional interest is the rather extended growth of crystals, which had nucleated at the edge of the film, over the support plate. In these cases, the crystal form was maintained even though the growth took place over a polycrystalline substrate. This means that the crystal form was determined in the embryonic

FIG. 4. Sodium chloride crystals on polycrystalline silver. Electropolished and etched. Note variation of density of crystallization from grain to grain and between twinned portions of grains. 150×

stage of development and subsequent growth was unaffected by the plate. This point will be amplified in the discussion.

On the silver film deposited on mica, the alignment was strict. The sodium chloride crystals assumed triangular pyramidal shapes. The results are presented in Fig. 3.

Striking effects were noted with polycrystalline silver. Practically complete epitaxy occurred for silver prepared in any of the following manners: electropolished and etched; electropolished but not etched; neither polished nor etched. In all cases, the surface was cleaned and heat-treated at 700°C for one hour prior to the application of the sodium chloride solution. The nature of the oriented overgrowth was the same for each of the three methods of surface preparation.

Characteristic results for a polished and etched specimen are shown in Fig. 4. The grain and twin boundaries may be seen in the field. The twinned regions are the generally banded areas within a grain. Each part of the grain which is twinned with the same orientation etches in the same way. Twinning is rather general over the whole area. Silver is well known to have a propensity for this behavior.

Those areas within a grain which etch in the same way are known to have the same orientation. Inspection of Fig. 4 reveals that the sodium chloride crystals had the same orientation over those parts of the grain which had the same appearance after being etched. In the dark areas of the large grain in the center of the figure, all the crystals had the same orientation. The white areas, which were devoid of crystals, were twinned regions of that grain.

The crystals on neighboring grains were oriented in various directions but on each grain one orientation was

preferred. On the twinned region of some of these grains no crystallization was observed. On others the crystals on the twins were all rotated through an angle with respect to those on the rest of the grain. Several examples of these effects may be seen in Fig. 4.

On polished but unetched silver similar behavior was noted. The crystals of sodium chloride clearly indicated the location and size of the grains. The similarity of the appearance of the sodium chloride crystals shown in the center of Fig. 5 and those in Fig. 1 suggest that the orientation of the substrate in the two cases was the same. Presumably the (100) plane of the grain in Fig. 5 lay in the surface of the plate.

It might have been anticipated that the discontinuities at the grain boundaries would provide places of ready nucleation. However, the grain boundaries, whether the silver was etched or not, were not preferred sites for crystallization. There was no evidence for such preference on any of the surfaces examined.

The density of crystallization varied from grain to grain. On some grains, in fact, no crystallization took place. The variation in density of crystallization was applied in one case to reveal the grain structure. A strongly alcoholic solution dried rapidly on a polished

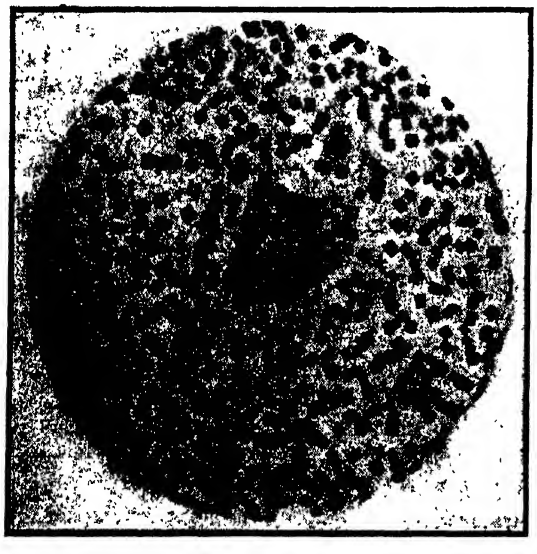


FIG. 5. Sodium chloride crystals on polycrystalline silver sheet. Electropolished and unetched. 150×

surface in a current of warm air was used to produce crystals ~1 micron in diameter. The variations in the density of crystallization produced sufficient contrast to outline the grains (Fig. 6). The same area was then rinsed and etched to delineate the grains in the usual manner. The result of this is shown in Fig. 7. The faithfulness of the reproduction by crystallization is apparent through a comparison of Figs. 6 and 7.

Another observation of interest was the growth of crystals across grain or twin boundaries without modification. From this it appeared that the form of the crystal was determined in the early stages of growth. Many examples of crystals which had nucleated on one side of a boundary and grown across may be found in Figs. 4 and 5. Additional illustrations were furnished when caesium chloride instead of sodium chloride was crystallized on the surface. The crystals grew farther into adjoining grains (Fig. 8) than did sodium chloride crystals. These results suggest that the influence of the substrate was exerted only in the embryonic stages of crystallization. Once the crystal had nucleated, subsequent growth was not affected by the nature of the

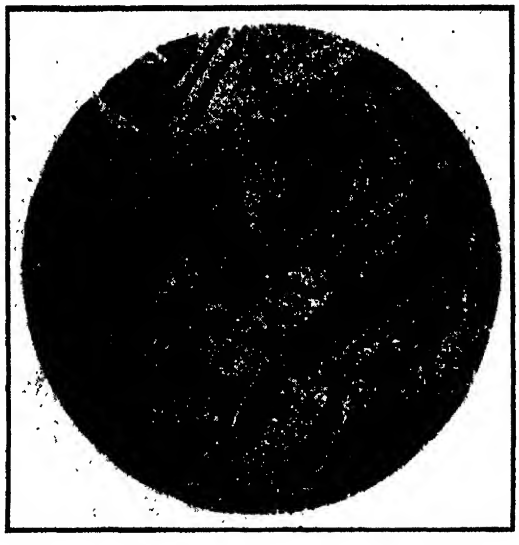


FIG. 6. Grain structure of polished polycrystalline silver revealed by variation of density of crystallization on the surface. 150×

underlying material. Further support to this view was given by the growth of crystals from the edge of the thin film shown in Fig. 2.

A silver surface, once it had been prepared so that it was capable of orienting sodium chloride, retained this capacity for months. There was no evidence of any aging effects.

DISCUSSION

There are four general features of the results which are of interest: (1) Pronounced epitaxy was found to exist between a metallic surface and an ionic crystal; (2) each area of the substrate of a given orientation was characterized by a particular orientation of the overgrowth; (3) the density of crystallization was dependent upon the orientation of the grain; (4) the crystals of sodium chloride often grew across grain and twin boundaries without modification. Each of these points is discussed in the following.

At the outset, it is clear that the third condition suggested by Royer,[1] requiring that the overgrowth and the substrate have the same type of binding, must be either modified or abandoned. Here we have strong evidence of epitaxy between a substrate with metallic binding and

an ionic crystal. The only question which might arise in interpretation is whether the salt was oriented directly by the substrate or by a compound on the surface. In particular the presence of an oxide layer would be suspected. Since each grain of silver in the polycrystalline specimens produced either a unique alignment or none, it follows that if an oxide layer were present it bore a definite crystallographic relationship to each grain of the substrate. While there are many such examples of oriented growth of oxides on metals, none have been reported for silver. Actually silver oxide is quite unstable

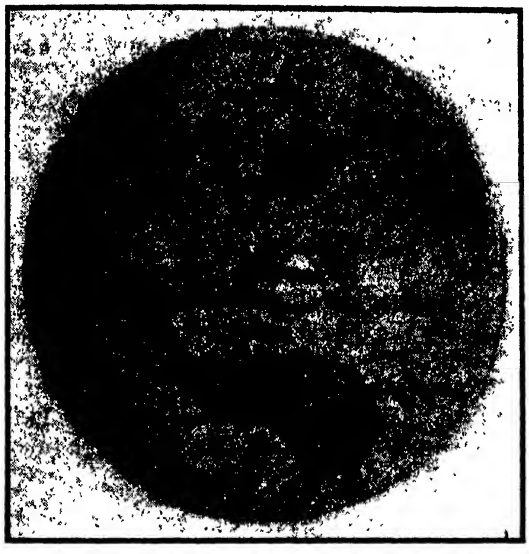


FIG. 7. Same area of polycrystalline silver sheet shown in Fig. 6 after etching. 150×

above 200°C and its rate of formation is exceedingly low at lower temperatures. As a result of an extensive investigation of the stability of silver oxide (Ag_2O) and the adsorption of oxygen on silver, Benton and Drake[13] concluded that silver acted upon by oxygen at room temperature becomes covered with a monomolecular layer. The rate of oxide formation would be negligibly small. From photoelectric work function measurements on silver, Brewer[14] found that an oxidized surface could be restored to a value characteristic of clean silver by heating at 600°C for several hours under vacuum. It will be recalled that the polycrystalline silver used in the present work was baked in a vacuum at 700°C for one hour and cooled in the furnace; and that the silver films were prepared by vacuum distillation. Consequently, while an adsorbed layer of oxygen probably existed on the surface, it is unlikely that an oxide film was present.

The results for polycrystalline metal and for oriented films indicated that the fit of the two lattices in contact need not be close for epitaxy to occur. On polycrystalline silver some form of alignment was produced on almost every grain. The universality of the effect in itself suggests that closeness of fit cannot be a strict requirement. Quantitative justification of this conclusion may be found in the results for oriented films.

When crystallized on a (100) plane of silver, the [100] direction of sodium chloride was directed along the [110] direction of the metal. The distance between adjacent atoms along the [100] axis of sodium chloride is 2.81A, and that between consecutive silver atoms along the [110] is 2.88A. Thus, with this arrangement the difference in atomic spacings 2.5 percent. As pointed out by G. P. Thomson[8] this relative orientation is the one of closest fit.

For (111) planes in contact, which was the situation for the film oriented by mica, the differences in atomic spacings of the sodium chloride and the silver lattice amounted to 73 percent. The existence of epitaxy between silver and sodium chloride, therefore, does not depend markedly on closeness of fit of the two lattices. Sloat and Menziers[3] obtained oriented growth on silver for only those salts whose lattice parameters differed from those of the metal by less than one percent; and, they concluded that the lattices must be commensurate

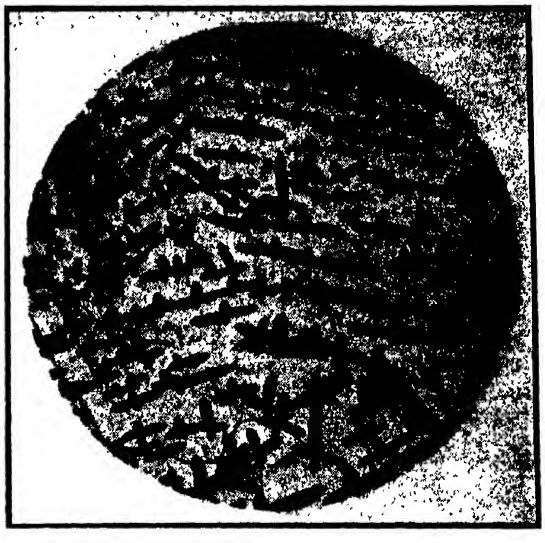


FIG. 8. Epitaxy of caesium chloride on polycrystalline silver. Note extensive growth of crystals across grain and twin boundaries without modification. 150×

within that tolerance. However, the present results show a much greater difference in lattice dimensions can be tolerated.

The observation that the density of crystallization depended upon the orientation of the substrate might have been anticipated. For electrodeposited metals, Finch and Sun[15] found that the adherence between the coat and the underlying surface was greatest when strong epitaxy existed between the two metals. Royer[1] noted that when a mixture of glass, mica, gypsum,

[13] A. F. Benton and L. C. Drake, J. Am. Chem. Soc. 56, 255 (1934).
[14] A. K. Brewer, J. Am. Chem. Soc. 56, 1909 (1934).
[15] G. I. Finch and C. H. Sun, Trans. Farady Soc. 32, 852 (1936).

calcite, talc, and orpiment were placed in a solution of ammonium iodide, the crystals always appeared first on the mica. The presence of epitaxy implies strong interatomic forces between the crystallizing atoms and those of the substrate. Therefore, a more rapid nucleation would be expected to take place on those surfaces favoring oriented growth. It is of interest to note that while contact of a crystal of a solute with a solution is the only means of completely preventing supersaturation, a surface which can show epitaxy with the solute may prevent a degree of supersaturation otherwise possible.[16]

For the silver film prepared on rock salt it was possible that nuclei were mechanically retained in the film after its removal from the salt. Such nuclei would serve as centers of crystallization when the film was tested for epitaxy. The growth, in this event, would have been parallel to the original rock salt lattice. Even after thorough rinsing a few crystals had this orientation, but practically all were rotated 45° with respect to this direction. Thus, the observed effects were undoubtedly characteristic of the silver film.

On the film with the (111) plane produced on mica, the observed results must be attributed to the silver, since Royer[10] showed that sodium chloride is not oriented directly by mica.

That the orientations of the crystals of the overgrowth were determined in the embryonic stages of crystallization was indicated by two observations: (1) the growth of crystals from the edges of an oriented film, and (2) the growth of crystals across grain or twin boundaries. Thus, the crystals often extended distances of ~0.01 mm, or several thousand atomic diameters, over areas of the substrate having an orientation different from that at the point of initial formation of the crystal. The fact that in many instances, a major portion of the crystal was resting on a surface which favored another orientation produced no modification of the overgrowth. More examples of this effect were found also with caesium chloride crystals on polycrystalline silver (Fig. 8). This situation could prevail only if the orientation of the crystal were determined on an area of much smaller dimensions than those of the final crystal. From the extent of growth across boundaries, a rough upper limit may be assigned to the size of the embryo. In many cases the crystallites were nearly centered on boundaries. Since the crystal diameters were about 0.01 mm, the embryo was certainly less than one-tenth of this value. Indeed, as pointed out by G. P. Thomson[8] and van der Merwe,[9] it is difficult to understand how the embryo could be larger than a few atomic diameters.

The epitaxy of sodium chloride provided a ready means of testing the degree of orientation of thin films without resorting to electron diffraction. Electron diffraction measurements are required to establish the crystallographic relationships of the overgrowth to the substrate, but once these are known, a rinse in sodium chloride solution suffices as a check. The rinsing procedure gives a representation of the orientation of the entire film, while the usual electron diffraction arrangement permits the examination of only a relatively small area. It must be emphasized however, that observations of epitaxy alone cannot be used to determine the orientation of the substrate, but can be used as a check after the expected orientation is known.

ACKNOWLEDGMENTS

The author is indebted to Dr. William Rubinson for a review of the literature and for many helpful criticisms; to Dr. David Gurinsky for constructive discussions; to Messrs. O. Kammerer and A. Cendrowski for the photomicrography; and to Mr. N. Bylock for technical assistance. All are members of the staff of the Brookhaven National Laboratory.

[16] This was pointed out by Dr. W. Rubinson. Private communication

Letters to the Editor

Note on Stability of Electron Flow in the Presence of Positive Ions

J. R. PIERCE
Bell Telephone Laboratories, Murray Hill, New Jersey
July 17, 1950

THE writer once calculated what he believed to be the limiting stable current in parallel electron flow normal to parallel conducting bounding planes (as, grids).[1] The conditions assumed were the following: Two parallel grids are spaced L cm apart. A current i_0 amp./cm^2 of electrons with a velocity u_0 specified by the accelerating voltage V_0 is injected through one. In the space between the grids, a positive charge density of immobile ions is assumed, such that if the electrons retain their velocity of injection the net space charge between the grids will be zero. The question is, for how large a value of i_0 will the flow be stable? The answer was sought by considering a small sinusoidal perturbation of the flow of frequency ω. It was found that if

$$\frac{\omega_0 L}{u_0} = \pi + \alpha \tag{1}$$

where α is a positive quantity small compared with π, the boundary conditions are satisfied for a frequency

$$\begin{aligned} \omega &= -j\pi u_0 \alpha / 4L \\ j\omega &= \pi u_0 \alpha / 4L. \end{aligned} \tag{2}$$

Here ω_0 is the electron plasma frequency.

$$\omega_0^2 = \frac{(e/m)\rho_0}{\epsilon} \tag{3}$$

where e/m is the charge-to-mass ratio of the electron, ρ_0 is the electronic charge density and ϵ is the permittivity of space. Thus, for $\omega_0 L/u_0$ a little greater than π, perturbations in the flow grow exponentially with time; the flow is unstable.

The situation analyzed was a little unrealistic in that the ions were assumed not to move. Under these circumstances, one obtains an expression for the phase constants of the space-charge waves

$$\beta = \frac{\omega}{u_0} \pm \frac{\omega_0}{u_0}. \tag{4}$$

If we allow motion of the ions, we find[2]

$$\beta = \frac{\omega}{u_0} \pm \frac{\omega_0}{u_0}\left(1 - \left(\frac{\omega_i}{\omega}\right)^2\right)^{-\frac{1}{2}}. \tag{5}$$

Here ω_i is the ion plasma frequency

$$\omega_i^2 = \frac{(q/m)\rho_i}{\epsilon} \tag{6}$$

where q/m is the charge-to-mass ratio of the ions and ρ_i is the ion charge density (assumed to be equal and opposite to the electron charge density).

We can use all the relations developed in reference 1 if we there replace ω_0 by ω_1, where

$$\omega_1 = \omega_0[1 - (\omega_i/\omega)^2]^{-\frac{1}{2}}. \tag{7}$$

Suppose, for instance, that we let

$$\frac{\omega_1 L}{u_0} = \pi + \alpha. \tag{8}$$

We now assume that α is small compared with π, so that ω is given in terms of α by (2). Then, using (7) and (8), we obtain

$$\alpha^2\left(\frac{\omega_0 L}{u_0}\right)^2 = (\pi+\alpha)^2\left[\alpha^2 - \left(\frac{\omega_i L}{u_0}\right)^2\left(\frac{4}{\pi}\right)^2\right]. \tag{9}$$

We have already assumed that $\alpha \ll \pi$, so α will be neglected in comparison with π in the parentheses. This gives for α

$$\alpha = \pm \frac{4(\omega_i L/u_0)}{[(\omega_0 L/u_0)^2 - \pi^2]^{\frac{1}{2}}}. \tag{10}$$

We note that for

$$\frac{\omega_0 L}{u_0} < \pi \tag{11}$$

α will be imaginary, the frequency will be real and the flow will be stable. For

$$\frac{\omega_0 L}{u_0} > \pi, \tag{12}$$

α will be real and positive for one root and hence $j\omega$ will be a real positive quantity and the beam will be unstable.

This argument does not apply near

$$\frac{\omega_0 L}{u_0} = \pi \tag{13}$$

since α will be large in this vicinity. We note, however, that for equal electron and ion charge densities

$$\frac{(\omega_i L/u_0)}{(\omega_0 L/u_0)} = \left(\frac{\text{mass of electron}}{\text{mass of ion}}\right)^{\frac{1}{2}}. \tag{14}$$

Thus, for massive ions, α will still be small compared with π for $(\omega_0 L/u_0)$ fairly near to π.

The conclusion then seems to be that even when the positive charge consists of massive ions rather than a fixed charge, there is a change from a stable to an unstable condition near (13), which represents the value of the electron plasma frequency ω_0 for which the flow becomes unstable for a fixed positive charge. It is of course possible that some instability for a lower value of ω_0 has been overlooked.

This note is to be construed merely as an amplification and clarification of work published earlier. Experiments seem generally to indicate that electron flow becomes unstable for currents nearer to those calculated assuming no positive ions present than to the larger currents calculated for the presence of ions.

[1] J. R. Pierce, J. App. Phys. 15, 721–726 (October, 1944).
[2] J. R. Pierce, J. App. Phys. 19, 231 (1948).

A Modification of the Radiofrequency Mass Spectrometer

A. F. HENSON
Imperial Chemical Industries Ltd., Butterwick Research Laboratories, The Frythe, Welwyn, Herts, England
April 26, 1950

THE resolution of a single stage of the radiofrequency mass spectrometer described by Bennett[1] is poor, and in order to obtain sufficient resolution a three-stage instrument is used. A possible method of obtaining the necessary resolution without the use of so many grids in the tube would be to apply a square wave form alternating potential to a single stage, the two halves of which were separated by constant potential "velocity selection" section. The tube would consist of the ion source as in Bennett's tube, four grids and a collector plate. For the analysis of positive ions, fixed potentials (negative with respect to the ion source) would be applied to the four grids. Grids 2 and 3 would be connected together and to them would also be applied the square wave form consisting of single complete cycles (positive half followed by negative) separated by intervals. The spaces between grids, the square-wave frequency and the intervals between cycles, would be arranged so that an ion of the selected mass passing through grid 1 at the start of a negative half-cycle passes through grid 2 at the end of the negative half-cycle. During the interval between cycles it travels in the field-free space between grids 2 and 3, emerges through grid 3 at the start of the next positive half-cycle, and through grid 4 at the end of this positive

half-cycle. It then passes on to the collector plate to which a blocking potential is applied in order to collect only the most energetic ions. It is thus accelerated by the square-wave field for the whole of the time it is between grids 1 and 2 and between grids 3 and 4.

A heavier ion than the selected one travels too slowly to gain maximum energy in either half-cycle. A lighter ion can gain maximum energy from either half-cycle but not from both. Thus ions of the selected mass gain more energy than any others. Different masses would be selected by varying the square-wave frequency and interval length, keeping the ratio of period of one cycle to interval constant.

A solution for the motion of the ions in the square-wave field shows that the discrimination between ions of the selected mass m (atomic weight units) and the next heavier ion $m+1$ is much larger than between m and $m-1$. The controlling consideration is then the discrimination between m and $m-1$. In the design of such an instrument there are many variables available. The values chosen will depend on the use to which the instrument is to be put and will be a compromise between conflicting requirements. Calculations have been made illustrating the use of the instrument for work in the low mass range up to about mass 50, such as might be used for stable isotope tracer work with carbon nitrogen and oxygen. The steady potential on the grids was -200 v and the amplitude of the square wave 60 v. It was considered that an energy discrimination between neighboring masses of at least 2 ev would be necessary and that, from consideration of ease of production of the square wave form, its period should not be less than 2 μsec. Under these conditions the tube length is about 80 cm and the energy discrimination between m and $m-1$ varies between 9.6 ev for $m=13$ and 2.4 ev for $m=50$. If it were convenient to produce the square wave form with greater amplitude or shorter period, or if a smaller energy discrimination were found to be sufficient, the tube could be considerably shortened.

[1] W. H. Bennett, J. App. Phys. 21, 143 (1950).

Remark on an Aberration-Phenomenon in Electrostatic Lenses. "Kidney Effect"

E. Bas-taymaz
Swiss Federal Institute of Technology, Department of Applied Physics, Zürich, Switzerland
June 7, 1950

In a letter,[1] S. Harrison described an aberration phenomenon in electrostatic lenses, the so-called "Kidney Effect" and gave an explanation. We had observed the same phenomenon in a special electron gun structure a long time ago, not being able to give any explanation. We give here a short description of our observation.

The scheme of the electron gun is shown in Fig. 1. It consists of a tungsten-Bolzen-cathode[2] and a focusing system similar to the description given by R. R. Law. The surface of emission can be imaged sharply on a fluorescent screen. However, with this

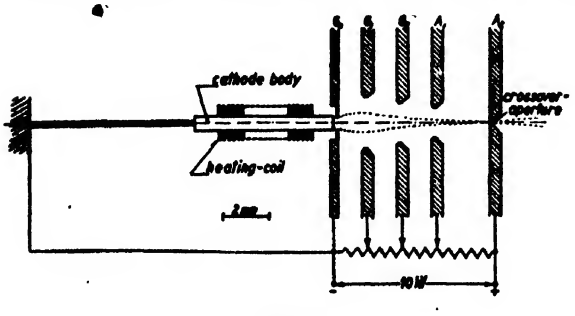


Fig. 1.

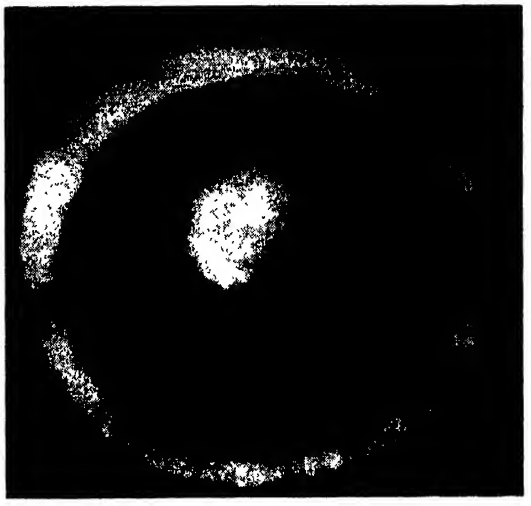


Fig. 2.

well-defined shape of the image produced by changing the strength of the lens by means of potentials of grids 1 and 2, a bright fluorescent spot occurred on the surroundings by further variation of the potentials, which grows to an illuminated circular area as shown in Fig. 2. The picture is the same as given by Harrison with only one difference that we have a greater symmetry resulting from the exact form of the electron gun. The picture, however, does not show exact symmetry and it shows how difficult it is to realize the condition of axial symmetry in electrostatic lenses.

The assumptions of Harrison make it possible to understand the phenomenon of the "Kidney Effect" in this special case, because we inserted *a priori* a little aperture in the focus (cross-over-aperture) which might cause the effect. Furthermore, we can see that this electron gun structure (Fig. 1) is showing a rather serious spherical aberration; a fact which is in accordance with our experience by which it has been ascertained that it is not possible to obtain the theoretical value of cross-over current density.

[1] S. Harrison, J. App. Phys. 20, 412 (1949).
[2] E. Baş-taymaz, Ph.D. thesis, E.T.H., Zurich, 1949 (to be published).

Dielectric Constants of Five Gases at 9400 Megacycles per Second*

James A. Murphy and Richard C. Raymond
Department of Physics, The Pennsylvania State College, State College, Pennsylvania
June 23, 1950

The dielectric constants of five gases have been measured at pressures up to one atmosphere by a resonant cavity method similar to that reported by Jen.[1] The cavity was a copper-walled cylinder with movable plunger and temperature control, and was operated in the TE01 circular mode. Measurements of the frequency shift and Q were taken from an oscillographic display of the effective resonance curve with the help of a coupled microwave circuit. Modulation power for measurement of frequency shifts was generated by a General Radio Type 1001A signal generator and could be varied in frequency from 5 kc/sec. to 50 Mc/sec. The minimum observable change in frequency was 0.25 Mc/sec.

The dielectric constants of the gases, as observed from the resonant frequency shift at a pressure of 730 mm mercury, are listed in Table I.

TABLE I. Dielectric constants of gases computed from observed resonant frequency shift at a pressure of 730 mm of mercury and wave-length 3.19 cm.

Gas	Dielectric constant $(\epsilon'-1)\times10^4$	Temperature
Acetone	17.1	75° C
Ammonia	6.8	26°
Carbon dioxide	1.6	26°
Carbon tetrachloride	3.9	85°
Water vapor	6.6	102°

Acetone and ammonia showed appreciable broadening of the resonance curves at the maximum pressures investigated. The loss factors were computed to be 2.7×10^{-4} and 4.7×10^{-4}, respectively, at the pressures and temperatures indicated in Table I.

The results are in general agreement with those reported by others[3-8] with the exception of CO_2. The CO_2 used here was prepared by sublimation of commercial dry ice, and may have included some water vapor.

* This paper is from a thesis by James A. Murphy submitted in partial fulfillment of the requirements for the degree of Master of Science at the Graduate School of The Pennsylvania State College.

[1] C. K. Jen, J. App. Phys. 19, 649–653 (July, 1948).
[2] C. M. Crain, Phys. Rev. 74, 691–693 (1948).
[3] L. G. Hector and D. L. Woernley, Phys. Rev. 69, 101 (1946).
[4] J. E. Walter and W. D. Hershberger, J. App. Phys. 17, 814–822 (1946).
[5] C. D. Hodgman, *Handbook of Chemistry and Physics* (Chemical Rubber Publishing Company, Cleveland, Ohio, 1937), p. 1451.
[6] J. A. Saxton, Rep. Phys. Soc., London, 215–238 (1947).

A Simplified Air Flow Method for the Determination of the Surface Area of Powders

GERARD KRAUS AND JOHN R. THIEM
Applied Science Research Laboratory, University of Cincinnati, Cincinnati, Ohio

IN a recent paper appearing in this Journal, F. A. Schwertz[1] reported a series of surface area measurements on sintered Pyrex diaphragms by means of a gas flow method similar to those of Rigden[2] and Arnell.[3] A particularly useful modification of this method which is quite rapid and requires only the simplest kind of equipment has been used by the writers to measure the surface areas of powders ranging from relatively coarse granules to colloidal carbon blacks.

The specific flow rate in moles/sec. per unit pressure difference under combined Poiseuille and slip flow is, in the notation of Schwertz (reference 1, Eq. (15)):

$$G=\frac{\epsilon^3AP_m}{k_v(1-\epsilon)^2S^2\mu LRT}+\frac{\epsilon^2\pi A}{k_d(1-\epsilon)SL(2\pi MRT)^{\frac{1}{2}}}\cdot\left(\frac{2}{f}-1\right). \quad (1)$$

Now consider a partially evacuated reservoir connected to the atmosphere through a plug of packed powder whose surface is to be measured. If the atmospheric pressure is P_0, the pressure inside the reservoir is $(P_0-\Delta P)$ and its volume V, it can easily be shown that

$$\frac{d(\Delta P)}{dt}=\frac{A\epsilon^3}{10V\mu L(1-\epsilon)^2S^2}(\Delta P)^2-\left[\frac{A\epsilon^3P_0}{5V\mu L(1-\epsilon)^2S^2}+\frac{A\epsilon^2Z}{(1-\epsilon)VSL}\left(\frac{\pi RT}{2M}\right)^{\frac{1}{2}}\right]\Delta P \quad (2)$$
$$=\beta(\Delta P)^2-\alpha\Delta P$$

where k_v has been taken as 5.0 and $(1/k_d)(2/f-1)$ has been replaced by Z. Dividing through by ΔP it follows that

$$\frac{d[\ln(\Delta P)]}{dt}=\beta\Delta P-\alpha \quad (3)$$

so that as ΔP approaches zero the slope of a plot of $\ln(\Delta P)$ *versus* t will be α.

TABLE I. Comparison of modified air flow method with established methods of surface area measurement.

	Surface area m^2/g		
	Air flow	Check	Method of check
Microscopic glass spheres	0.136	0.139	Microscopic count
Carbon black (Thermax)	7.4	7.69	Nitrogen adsorption[a]

[a] P. H. Emmett and T. De Witt, Ind. Eng. Chem., Anal. Ed. 13, 28 (1941).

Integration and rearrangement of (2) gives

$$e^{\alpha t}-1=\frac{\alpha}{\beta}\left[\frac{e^{\alpha t}}{(\Delta P)_0}-\frac{1}{\Delta P}\right] \quad (4)$$

where $(\Delta P)_0$ is the initial pressure difference. A plot of $(e^{\alpha t}-1)$ against the quantity in brackets yields a straight line of slope α/β. By definition

$$\beta=\frac{A\epsilon^3}{10V\mu L(1-\epsilon)^2S^2}$$

and hence

$$S=\frac{\epsilon}{1-\epsilon}\left(\frac{A\epsilon}{10V\mu L\beta}\right)^{\frac{1}{2}}. \quad (5)$$

The use of Eq. (1) under varying pressure difference involves an approximation due to neglect of inertia forces; however, the error thus introduced is entirely negligible.

In practice it is most convenient to take two sets of pressure-time readings; one with P varying from 35 to 10 cm Hg and the other in the range of 2 to 0.5 cm Hg using an oil-filled manometer. The latter set of data is used for the $\log\Delta P$ *versus* time plot.

Table I gives two examples of results obtained by the above procedure. Both test plots gave excellent straight lines. The agreement of the surface areas with independent methods of measurement is seen to be entirely satisfactory. The glass microspheres were prepared in the laboratory. A count of 501 particles gave a normal distribution with a mean diameter of 14.21 microns and a standard deviation of 4.60 microns from which the value of S in Table I was calculated.

[1] F. A. Schwertz, J. App. Phys. 20, 1070 (1949).
[2] P. J. Rigden, J. Soc. Chem. Ind. 66, 130 (1947).
[3] J. C. Arnell, Can. J. Research 24A, 103 (1946); 25A, 191 (1947); 26A, 29 (1948).

The Pursuit Course

RALPH HOYT BACON
The Perkin-Elmer Corporation, Elmhurst, New York
June 23, 1950

SOME properties of the path of a particle traveling along a curve whose tangent always passes through a point target moving with uniform speed along a straight line have been mentioned by Yuan.[1] Perhaps a fuller description of this curve, called the pursuit course, might be of interest.

Let the path of the target be the y axis, and let the x axis be tangent to the path of the seeker (or pursuer). Then, using Yuan's notation, let

$$\left.\begin{aligned} v_S&=Kv_T\\ dS&=Kdy_T\end{aligned}\right\}, \quad (1)$$

where dS is an element of arc of the pursuit course. For the slope of the tangent of course, at any point, we have

$$\tan\theta=dy_S/dx_S=(y_S-y_T)/x_S.$$

Therefore,

$$\frac{d^2y_S}{dx_S^2}=-\frac{1}{x_S}\frac{dy_T}{dx_S}=-\frac{1}{Kx_S}\frac{dS}{dx_S}=\frac{[1+(dy_S/dx_S)^2]^{\frac{1}{2}}}{Kx_S}.$$

Transposing,

$$\frac{d(dy_S/dx_S)}{[1+(dy_S/dx_S)^2]^{\frac{1}{2}}}=\frac{dx_S}{Kx_S}.$$

Integrating,

$$2\frac{dy_S}{dx_S}=\left(\frac{x_S}{x_0}\right)^{1/K}-\left(\frac{x_S}{x_0}\right)^{-1/K}, \tag{2}$$

where x_0, the constant of integration, is the abscissa of the point of contact of the pursuit curve with the x axis.

Integrating again,

$$\left.\begin{aligned} y_S&=\frac{Kx_S}{2}\left[\frac{1}{1+K}\left(\frac{x_S}{x_0}\right)^{1/K}-\frac{1}{K-1}\left(\frac{x_S}{x_0}\right)^{-1/K}\right]+\frac{Kx_0}{K^2-1} \quad (K\neq 1)\\ y_S&=\frac{x_S^2-x_0^2[1+2\log(x_S/x_0)]}{4x_0} \quad (K=1)\end{aligned}\right\}. \tag{3}$$

For the length of path traveled by the seeker,

$$dS=-\left[1+\left(\frac{dy_S}{dx_0}\right)^2\right]^{\frac{1}{2}}dx_S=-\frac{1}{2}\left[\left(\frac{x_S}{x_0}\right)^{1/K}+\left(\frac{x_S}{x_0}\right)^{-1/K}\right]dx_S.$$

Therefore,

$$\left.\begin{aligned} S&=\frac{K^2x_0}{K^2-1}-\frac{Kx_S}{2}\left[\frac{1}{K+1}\left(\frac{x_S}{x_0}\right)^{1/K}+\frac{1}{K-1}\left(\frac{x_S}{x_0}\right)^{-1/K}\right] \quad (K\neq 1)\\ S&=\frac{x_0^2[1-2\log(x_S/x_0)]-x_S^2}{4x_0} \quad (K=1)\end{aligned}\right\}, \tag{4}$$

where S is measured, plus or minus, from the point of tangency, the point at which $x_S=x_0$.

Combining Eqs. (3) and (4):

$$\left.\begin{aligned} S-y_S&=\frac{K}{K+1}\left[x_0-x_S\left(\frac{x_S}{x_0}\right)^{1/K}\right]\\ &=(x_0^2-x_S^2)/2x_0 \text{ when } K=1\end{aligned}\right\}. \tag{5}$$

Let the distance between the seeker and the target be denoted by P. Then

$$P^2=x_S^2+(y_S-y_T)^2=x_S^2\left[1+\left(\frac{y_S-y_T}{x_S}\right)^2\right]$$

$$=x_S^2\left[1+\left(\frac{dy_S}{dx_S}\right)^2\right]=\frac{x_S^2}{4}\left[\left(\frac{x_S}{x_0}\right)^{1/K}+\left(\frac{x_S}{x_0}\right)^{-1/K}\right]^2.$$

Therefore,

$$P=\frac{x_S}{2}\left[\left(\frac{x_S}{x_0}\right)^{1/K}+\left(\frac{x_S}{x_0}\right)^{-1/K}\right]. \tag{6}$$

Finally, for the radius of curvature of the course,

$$R=\frac{[1+(dy_S/dx_S)^2]^{\frac{1}{2}}}{d^2y_S/dx_S^2}=\frac{Kx_S}{4}\left[\left(\frac{x_S}{x_0}\right)^{1/K}+\left(\frac{x_S}{x_0}\right)^{-1/K}\right]^2. \tag{7}$$

Now, as x_S approaches zero,

$$y_S\to\frac{Kx_0}{K^2-1}-\frac{Kx_0^{1/K}}{2(K-1)}x_S^{(K-1)/K}, \text{ when } K\neq 1;$$

$$\to\frac{Kx_0}{K^2-1} \text{ when } K>1;$$

$$\to\infty \text{ when } K=1 \text{ or when } K<1.$$

$$S\to\frac{K^2x_0}{K^2-1} \text{ when } K>1;$$

$$\to\infty \text{ when } K=1 \text{ or when } K<1.$$

$$S-y_S\to\frac{Kx_0}{K+1}.$$

$$P\to 0 \text{ when } K>1;$$

$$P\to\frac{x_0}{2} \text{ when } K=1;$$

$$P\to\infty \text{ when } K<1.$$

$$R\to 0 \text{ when } K>2;$$

$$\to\frac{x_0}{2} \text{ when } K=2;$$

$$\to\infty \text{ when } K<2.$$

The fact that R approaches zero if K be greater than 2 is well know to every student of the subject.

All of this is an interesting exercise in mathematics; the limitations of its usefulness as a tactical maneuver are pointed by Yuan,[1] who suggests an alternative course.

[1] Luke Chia-Liu Yuan, J. App. Phys. 19, 1122 (1948).

Some Observations on the Back Heating of Magnetron Cathodes

R. T. Young, Jr., L. W. Holmboe, and W. E. Waters, Jr.
National Bureau of Standards, Washington, D. C.
May 24, 1950

In the course of work on *CW K*-band magnetrons, we have observed some cathode back heating effects which we consider to be of sufficient interest to be worth recording. The magnetrons were of the interdigital type.[1] The pertinent dimensions of a typical tube are: anode radius—0.16 cm, cathode radius—0.11 cm, active cathode length—0.12 cm, number of teeth—20. The fundamental, first-, second-, and third-order cavity modes for this tube were at 2.9, 1.8, 1.2, and 0.9 cm, respectively. The tubes were designed to operate at 2000–3000 volts, 4000–5000 gauss and 30 ma, some in the second-order mode and some in the first-order mode. A catephoretically deposited thoria on molybdenum cathode was employed.

The experimental results on all tubes were of the same general nature, namely: (a) The tube operated very inefficiently in the desired cavity modes. This type of operation was possible only at low currents, and only under very critical conditions of voltage, magnetic field, and particularly cathode temperature. (b) When either the cathode temperature or the anode voltage was raised, the tube would break discontinuously into an undesired state of oscillation characterized by large back bombardment of the cathode, a marked increase in anode current, and a low power broad-band noise output. The back bombardment power was estimated to be at least 10 watts since it was enough to keep the cathode operating without heater power.

The observed noise output was centered at 1.8 and 1.2 cm, the wave-lengths of the first- and second-order cavity modes, with the stronger signal at 1.8 cm. The frequency separation between the half-power points corresponded roughly to the Q of the cavity. These observations were made on a spectrum analyzer; the power was too small to detect with a thermistor bridge sensitive to 25 microwatts. A signal in the fundamental cavity mode would not be detected because the wave guide was beyond cut-off for the wave-length of that mode. To avoid the frequency limitation of the wave-guide output, a special tube was constructed with a coaxial probe electrostatically coupled to the tooth structure and oriented so as to couple to any degenerate modes. No significant differences in the output could be detected. Our interpretation of these observations is that the cavity acted as a filter for noise generated in the interaction space.

To corroborate the fact that the cavity was not supporting this undesired oscillation, a smooth anode magnetron with identical end space and pole piece configuration, but with a solid cylindrical anode replacing the tooth structure, was tested. The same back bombardment phenomena were observed. A careful search was made for radiation coming out of the cathode leads using a horn feeding X-band wave guide, but none could be detected.

Plots of anode voltage *vs.* anode current for the above oscillations were made at magnetic fields from 1200 to 6200 gauss for both the interdigital and smooth anode tubes. They were all similar to the *V-I* curves of a normal cavity magnetron. The curves of voltage *vs.* magnetic field also were similar to the *V-B* curves of a normal cavity magnetron. That is, they were straight lines, approximately tangent to the cut-off parabola. These results held for tubes with a variety of end space configurations, magnetic field distributions, and ratios of anode-cathode radii.

It would appear that a state of oscillation is involved which is primarily a function of the space charge conditions in the interaction space. Since we have not identified any frequency associated with the oscillation, it is difficult to draw any conclusions as to its nature. One supposition is that there is a generation of noise alone. However, one cannot rule out the possibility that the observed signals are noise side bands of a coherent signal. From the linearity between V and B, one concludes that the condition for the oscillation depends upon the angular velocity of the electrons.

The phenomenon of back heating in smooth anode magnetrons has been noted by several authors[2-4] and has generally been attributed to some type of oscillation. There is no completely satisfying explanation available. Using a rising sun design, Columbia Radiation Laboratory[5] has obtained up to 30 watts from K-band magnetrons under roughly the same operating conditions as ours. Their reports briefly mention some back heating effects which probably are due to the same type of oscillation as reported here, but these effects do not prevent normal oscillation in their tubes.

It appears that in the magnetrons we worked with this undesired oscillation is sufficiently pronounced to prevent normal operation of the tube. An understanding of this type of oscillation may be of fundamental importance in the design of low voltage, short wave-length magnetrons.

[1] F. H. Crawford and M. D. Hare, Proc. I.R.E. 35, 4, 361-369 (1947).
[2] E. G. Linder, Proc. I.R.E. 24, 633-653 (April, 1936).
[3] E. G. Linder, Proc. I.R.E. 26, 346-371 (March, 1938).
[4] E. C. S. Megaw, Nature 132, 854 (December, 1933).
[5] Columbia Radiation Laboratory Quarterly Progress Report, December 31, 1947 and 1948 reports, Signal Corps Contract W-36-039 sc-32003.

Some Magneto-Acoustic Effects in Nickel*

T. F. Rogers
Air Force Cambridge Research Laboratories, Cambridge, Massachusetts
AND
S. J. Johnson
Crystal Research Laboratories, Inc., Hartford, Connecticut
June 1, 1950

An investigation has been made of effects produced by a magnetic field upon the ultrasonic propagation properties of certain ferromagnetic substances, of which nickel is reported upon here. A rod of commercially pure—99.4 percent—"A" nickel 45.8 cm long and 1.25 cm in diameter employing a.c.-cut quartz crystal transmitting and receiving transducers, exhibits an "electrical length" of some 170 μsec. when acting as a propagation medium for a short-pulsed, 1-Mc sonic carrier frequency, with the transmitting crystal generating a transverse (shear) wave within the rod. After having been heated to approximately 1600°F, a red heat assuring demagnetization, and allowed to cool slowly to room temperature, 70°F, the rod was subjected to a longitudinal magnetic field throughout its entire length; changes in its physical length were measured directly, while transmission (delay) time and relative attenuation were measured by means of an oscilloscope reproduction of the received pulse, both as functions of the applied field strength.

Figure 1 indicates the change—a decrease—in transverse wave propagation time delay (electrical length) ΔT, relative to the original time delay T, as a function of the field intensity in oersteds. The observed changes in delay of approximately 10^3 times that expected from the magnetostrictive (Joule) effect in nickel can, apparently, only be explained as signifying a marked increase in the acoustic velocity of propagation within the rod while it is under the magnetic field stress.

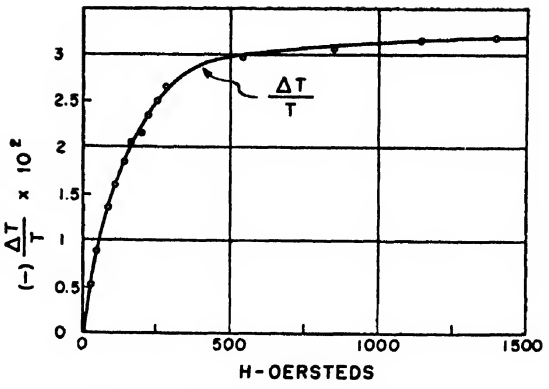


FIG. 1. Some magneto-acoustic effects in nickel.

The velocity of propagation of a transverse vibration is given as:[1]

$$V_t=\left[\frac{\mu}{\rho}\right]^{\frac{1}{2}}=\left[\frac{E}{2\rho(1+\sigma)}\right]^{\frac{1}{2}},$$

where μ is the modulus of rigidity, E is Young's modulus, ρ is the density, σ is Poisson's ratio, and $E=2\mu(1+\sigma)$. With the applied field, the change in density is considered negligible as a factor contributing to the velocity change, since it amounts to a factor of less than 10^{-4}, and, therefore, the increase in velocity must be attributed to a change in the value of the modulus of rigidity; the work of various groups[2] on the variation of E with magnetization (the so-called ΔE effect) would also appear to warrant this conclusion.

A large decrease in attenuation was also noted upon application of the field, with the received pulse increasing in amplitude some 10^2 times its normal value. Figure 2 presents a representative

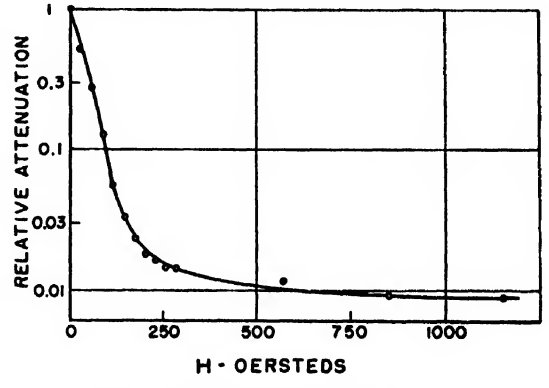


FIG. 2. Some magneto-acoustic effects in nickel.

change in attenuation throughout the transmission system as a function of H with the output pulse taken as unity height for $H=0$. It has been found that, for certain metallurgical conditions of the sample, irregularities are noted in the attenuation *vs.* H curve, with treatment of the rod immediately previous to test influencing the general march of values to a considerable extent. While possibly connected with a reorientation of the nickel crystalline structure, at this time, especially in view of the fact that ultrasonic attenuation effects in solid media are not altogether understood,[3] no adequate explanation of this decrease in attenuation presents itself.

It is planned that work on nickel, as well as other magnetic materials, will be continued and extended. These initial results are presented at this time for reasons of the natural interest in

such originally noted phenomena and in the expectation that this investigative technique will lend itself to a study of acoustic propagation in magnetic media and the structure of the media themselves.

* This work was supported by the Air Force Cambridge Research Laboratories, Cambridge, Massachusetts.

[1] L. Bergmann, *Ultrasonics* (John Wiley and Sons, Inc., New York, 1939), p. 162.

[2] E. Giebe and E. Blechschmidt, Ann. d. Physik LPZ (v) 11, 905 (1931); S. Siegal and S. L. Quimby, Phys. Rev. 49, 663 (1936); R. Kimura, Proc. Phys. Math. Soc. Japan 21, 786 (1939).

[3] C. Kittel, *Reports on Progress in Physics* (The Physical Society, London, 1948), Vol. XI (1946–47), p. 205.

Electron-Microradiography of Electrodeposited Metals

ROLF WEIL AND HAROLD J. READ
Division of Metallurgy, Pennsylvania State College, State College, Pennsylvania
August 2, 1950

IN the course of research on the structure of electrodeposited metals, a new method of examining specimens directly in the electron microscope, without the aid of replicas, was developed. This method is most valuable in the study of the initiation of plating, revealing some features not seen when conventional techniques are employed. As annealed electroplated metals are essentially the same as annealed metals prepared from the melt, the method described here should also be a useful tool in the study of other metallurgical phenomena.

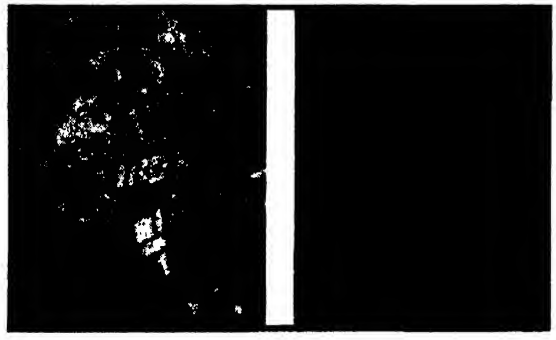


FIG. 1. (Upper left.) Electron-microradiography of electrodeposited nickel on copper. 5000 X.

FIG. 2. (Upper right.) Electron-micrograph from collodion replica of specimen shown in Fig. 1. 5000 X.

FIG. 3. (Lower left.) Electron-microradiograph of electrodeposited nickel on zinc. 5000 X.

FIG. 4. (Lower right.) Electron-micrograph from collodion replica of specimen shown in Fig. 3. 5000 X.

To date, nickel films plated on copper and zinc have been examined. The method of preparing the samples was as follows: The starting cathode was cleaned by the usual procedures for electroplating, and a deposit approximately 5×10^{-4} cm thick laid down. A piece the size of an electron-microscope screen was cut from the plated cathode, and placed in a suitable solvent to dissolve the basis-metal. For nickel-plated copper, a solution of 500 g/l CrO_3 and 50 g/l H_2SO_4 dissolved the copper, leaving the nickel. A solution of five percent by volume of concentrated HCl in water served for dissolving zinc basis-metals. When the original cathode was dissolved, the nickel film was washed and placed on a screen, ready for examination.

The accompanying photographs (Figs. 1–4) show the structures of nickel films deposited on copper and zinc respectively as seen by electron-microradiography and by the collodion replica techniques. The films were deposited under the same plating conditions, thus illustrating the effect of the basis-metal. The direct method resolved the structure far more distinctly than collodion replicas.

For convenience in referring to pictures made by the above technique, we should like to suggest the term "electron-microradiography" so as to distinguish conveniently such pictures from those obtained by the replica technique.

The Limiting Hydrostatic Tension of Water Near 0°C

J. C. FISHER
General Electric Research Laboratory, Schenectady, New York
August 3, 1950

RECENTLY L. J. Briggs has determined the limiting negative pressure in water as a function of temperature.[1] Above 5°C his values are in the neighborhood of 270 negative bars, about 20 percent of the theoretical value calculated from nucleation theory.[2] Below 5°C the limiting tension was observed to drop rapidly with decreasing temperature to a value near zero just above 0°C.

The rapid decrease of limiting tension can be explained in terms of the nucleation of ice under reduced pressure. Assume that ice nuclei form readily and grow when the solid phase is stable, and that ice nucleates the vapor phase. The assumption that ice nucleates water vapor is consistent with the lower solubility of gasses in ice and the tendency of gas bubbles to form at the surface of a growing crystal. Then near 0°C the limiting tension that water will withstand is that required to raise the freezing point to the testing temperature.

A negative pressure of 300 bars raises the freezing point of water 2.7°C, in fair agreement with Briggs' observations. A possible check of the ice nucleation hypothesis would be to determine the supercooling limit of the water at atmospheric pressure. If appreciable supercooling is observed, it is unlikely that ice nuclei form as readily as would be required to account for the decrease in limiting tension.

[1] L. J. Briggs, J. App. Phys. 21, 721 (1950).
[2] J. C. Fisher, J. App. Phys. 19, 1062 (1948).

Local Order in Solid Cu_3Au

R. A. ORIANI
Research Laboratory, General Electric Company, Schenectady, New York
June 5, 1950

RECENT x-ray data by Cowley[1] on Cu_3Au at temperatures above the critical temperature of superlattice formation enable the degree of local order computed from x-ray measurements to be compared with that calculated from thermodynamic

activity coefficients measured by Weibke and Quadt.[3] Such a comparison is of great interest, for this is the first system, to the writer's knowledge, for which two independent sets of data at overlapping temperatures have been obtained from which local order may be computed by two different methods of analysis. This note briefly summarizes calculations of local order at the composition Cu_3Au at 500°C, over 100° higher than the critical temperature of superlattice formation.

1. *X-ray data.*—Cowley's short-range order parameters for the first coordination shell at 405, 460, and 550°C were employed to interpolate the value of $\alpha = -0.142$ at 500°C, since the e.m.f. data for 500°C were used. In Cowley's terminology,

$$\alpha = 1 - (P_{l,m,n}/m_{Cu}), \tag{1}$$

where $P_{l,m,n}$ = probability that the atom with coordinates l, m, n with respect to an Au atom should be a Cu atom; m_{Cu} = proportion of Cu atoms = 0.75 for Cu_3Au. For complete randomness, $\alpha = 0$ so that $P = m_{Cu}$. For perfect order, $\alpha = 1$ for even-numbered shells, and $\alpha = -\frac{1}{3}$ for odd-numbered shells. The experimental value of α, and hence the degree of local order, depends only on the precision of the experimental measurements and on the adequacy of the three-dimensional Fourier analysis of the diffuse background scattering of x-rays. From $\alpha = -0.142$, one easily computes that (a) for Cu at the origin, the average number of Cu atoms in the first coordination shell is 8.55, and of Au atoms, 3.45; (b) for Au at the origin, the average number of Cu atoms in the first coordination shell is 10.3, and of Au atoms, 1.7. Hence, x = the ratio of the number of Au–Au bonds to the total number of bonds = 0.0354, whereas for random distribution $x = 0.0625$. Hence, there are more bonds between dissimilar atoms than the number expected from a random distribution of atoms; the percent reduction of Au–Au bonds from that characteristic of a fully disordered structure is 43.4 percent.

2. *Activity data:* A. *First method.*—From Weibke and Quadt's 500°C e.m.f. data, a smoothed curve of $\ln\gamma_{Cu}$ *vs.* mole fraction of Cu was constructed; the activity coefficients, γ, are referred to pure copper as the standard state. From this curve $\gamma_{Cu} = 0.79$ at 0.75 mole fraction of Cu was obtained. The activity coefficient for gold was calculated as 0.143 by graphical integration of the Duhem-Margules equation.

Takagi[3] has developed a relation between x, the fraction of Au–Au bonds, and ΔF°, the free energy change for the bond reaction

$$\text{Au–Au} + \text{Cu–Cu} \rightarrow 2(\text{Au–Cu}). \tag{2}$$

Birchenall[4,5] independently derived the special case applicable to Cu_3Au, i.e.,

$$\frac{(1-4x)^3}{2x(1+2x)} = 4e^{-\Delta F^\circ/RT}. \tag{3}$$

In order to compute ΔF°, we make use of regular solution theory,[6] relating the activity coefficient of A to the difference between the partial molar internal energy of vaporization of an ideal solution of A in B and that of the pure metal. Following Birchenall,[4] we neglect the volume change in going from ideal to regular solution, and assume as a first approximation that the atoms in the solution are distributed randomly in order to compute the partial molar internal energy of vaporization of the solution. Then,

$$\Delta F^\circ = \Delta\bar{F}_{Cu} + \Delta\bar{F}_{Au} \cong \Delta\bar{E}_{Cu} + \Delta\bar{E}_{Au} = RT(\ln\gamma_{Cu} + \ln\gamma_{Au}),$$

and by use of (3), a value of x is obtained. The whole calculation may be repeated, correcting for the randomness assumed by using this value of x; the second approximation yields ΔF° for reaction (2) = −820 cal./mole of bonds, and x = fraction of Au–Au bonds = 0.0453, representing a 27.5 percent reduction in the number of Au–Au bonds from the number characteristic of randomness. In order to evaluate the kind of variation introduced by an error in the experimental data, the first part of this calculation was repeated using a γ_{Cu} some 3 percent lower than the 0.79 employed above, since this is about the scatter of the data. The value of x thus computed is less than 2 percent larger than that obtained above.

B. *Second method.*—Lo-Ching Chang,[7] using Takagi's development has shown that if there is no volume change on mixing, and if the pairwise energies of the atoms are independent of composition, the heat of solution,

$$\Delta H = (x_{11} - m_1)(v_{11} + v_{22} - 2v_{12}), \tag{4}$$

where x_{11} = fraction of 1–1 bonds, m_1 = mole fraction of component 1, and v = energy of a pair of atoms. Transcribing this into our nomenclature, we have

$$\Delta H = (1/2)Z\Delta F^\circ(0.25 - x), \tag{5}$$

where Z = coordination number = 12. This method of treatment is very similar to that of part (A), since Eq. (5) also makes use of the regular solution approximation. If we use the value of $\Delta H = -1000$ cal./mole obtained from Weibke and Quadt's activity data, and solve (3) and (5) simultaneously, we obtain

$\Delta F^\circ = -816$ cal./mole for reaction (2)
$x = 0.0455$
27.2 percent reduction in the number of Au–Au bonds.

In view of the difficulties of the measurements and the approximations involved in the thermodynamic treatment, the agreement among the three methods is quite satisfactory, and lends confidence to the use of activity coefficients in this manner. A similar satisfactory comparison between x-ray and thermodynamic measurements of local order on the Ag–Au system, though at two different temperatures, has been made by B. L. Averbach reported by C. E. Birchenall.[8]

[1] J. M. Cowley, J. App. Phys. 21, 24 (1950).
[2] F. Weibke and U. v. Quadt, Zeits. f. Elektrochemie 45, 715 (1935).
[3] Y. Takagi, Physico-Math. Soc. Japan 23, No. 1, 44 (1941).
[4] C. E. Birchenall, A.I.M.E. Tech. Pub. No. 2169 (1947).
[5] C. E. Birchenall and C. H. Cheng, Metals Trans. 185, 428 (1949).
[6] J. H. Hildebrand, *The Solubility of Non-Electrolytes* (Reinhold Publishing Corporation, New York, 1936).
[7] Lo-Ching Chang (unpublished).
[8] C. E. Birchenall, Am. Soc. Metals, Cleveland, 159 (1950).

Formation of Annealing Twins during Grain Growth

R. L. FULLMAN
Research Laboratory, General Electric Company, Schenectady, New York

June 5, 1950

A DOMINANT feature of the microstructures of most annealed metals and alloys with a face-centered cubic crystal structure is the presence of many straight-sided crystals that have a twinned orientation relative to the grains in which they appear. Mathewson[1] proposed that these "annealing twins" form during recrystallization, by lateral growth of "twin faults" produced by the slip process. Carpenter and Tamura[2] found that annealing twins may form under conditions where it appears doubtful that slip could be of importance. They suggested that annealing twins form by accidents of growth when grain growth occurs. Burke[3] found that, during grain growth, annealing twins appear, one boundary at a time, in the corners of growing grains. During continued grain growth the distance between twin boundaries was not observed to change.

The interfacial free energy of the boundary between adjacent grains is a function of the orientation relationship between the grains.[4] During grain growth, changes occur in the identity of a grain's neighbors, and the interfacial free energies of the grain boundaries also change. It is suggested that an annealing twin forms in the corner of a growing grain when the orientation relationships between the twin and the grain's neighbors lead to a smaller total interfacial free energy than do the orientation rela-

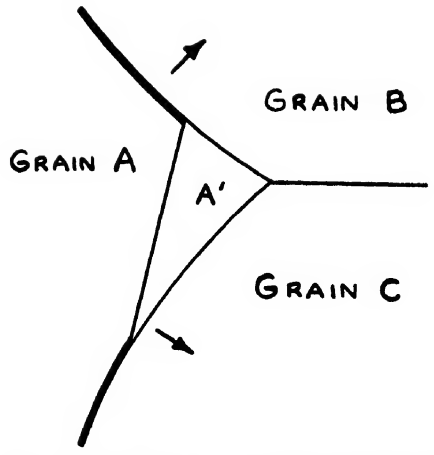


FIG. 1. Schematic representation of the changes in grain boundary free energy associated with the formation of a twin, A', in an infinitesimal region in the corner of a growing grain, A. Grain boundaries with high interfacial free energies are represented by wide lines.

tionships between the grain itself and its neighbors.* A twin boundary then forms in a corner of the grain, and the grain continues to grow as a twin of the original crystal. This could occur at various times during the growth of a grain, when the identity of one or more of the neighboring grains changes. The conditions required for a decrease in free energy may be clarified by reference to Fig. 1, which represents schematically a small region where atoms are being added to the corner of a growing grain, A, and removed from grains B and C. The small arrows indicate the directions of grain boundary migration, and the wide lines indicate boundaries with relatively high interfacial free energy. The free energy will be decreased by formation of a twin, A', in the corner of grain A, if

$$l_{A'B}\sigma_{A'B}+l_{A'C}\sigma_{A'C}+l_{AA'}\sigma_{AA'}<l_{A'B}\sigma_{AB}+l_{A'C}\sigma_{AC}, \quad (1)$$

where l and σ represent the areas and interfacial free energies, respectively, of the boundaries indicated by the subscripts. Annealing twins would form in grain corners during grain growth when the geometry and orientation relationships are such that inequality l is satisfied.

During grain growth the identity of a grain's neighbors may change by several mechanisms. Such changes may lead to the formation of new twins or the removal of existing twins. Details of the mechanism of formation and disappearance of annealing twins, and experimental evidence supporting the theory, will be given in forthcoming papers. Experiments designed to detect the difference in the interfacial free energies of grain boundary segments on opposite sides of twin boundaries newly formed in grain corners indicate that the free energy actually is smaller toward the grain corners. The dependence of the abundance of annealing twins in copper on the preferred orientation and grain size was found to be in agreement with predictions made on the basis of the proposed theory of twin formation and disappearance. The ratio of the interfacial free energy of twin boundaries to that of grain boundaries has been measured in copper and in aluminum. The values found were 0.036 ± 0.006 and 0.21 ± 0.05, respectively. Annealing twins rarely are observed in aluminum. The difference observed in the abundance of annealing twins in copper and aluminum is in accordance with the proposed theory and the difference found in the twin boundary to grain boundary interfacial free energy ratios.

This letter is an abstract of a dissertation presented to the Yale School of Engineering in partial fulfillment of the requirements for the degree of Doctor of Engineering. The investigation was supported by the United States AEC under Contract No. W-31-109-ENG-52.

[1] C. H. Mathewson, Trans. A.S.M. 32, 38 (1944).
[2] H. C. H. Carpenter and S. Tamura, Proc. Roy. Soc. 113A, 161 (1926).
[3] J. E. Burke (to be published).
[4] C. G. Dunn and F. Lionetti, Trans. A.I.M.E. 185, 125 (1949).

* This concept was suggested by J. C. Fisher, who has also assisted in the further development of the theory.

Journal
of
Applied Physics

Volume 21, Number 11 November, 1950

On Crazing of Linear High Polymers*

C. C. Hsiao and J. A. Sauer
The Pennsylvania State College, State College, Pennsylvania
(Received April 21, 1950)

The effect of various variables, such as type of stress, stress magnitude, duration of stressing, and environment on the initiation and development of crazing in linear polymers is discussed. The basic nature of crazing is investigated in some detail for polystyrene specimens by means of the light microscope, the electron microscope and the x-ray spectrometer. The results of these observations and their bearing on the fundamental group structure of polystyrene molecules is presented.

The relationship between crazing and orientation is discussed, as well as the effect of both of these factors on the mechanical properties. The experimental results are compared, wherever possible, with the previously reported data of other investigators. A short discussion is then given of a theory of crazing from the point of view of the molecular structure of the material.

INTRODUCTION

THE phenomenon of crazing of thermoplastic materials, especially polystyrene, has been under intensive investigation recently by many scientists.[1-5] One of the aims of this research is to discover the cause of crazing and to find methods of eliminating it, if possible. The inception and subsequent growth of crazing are undesirable features that tend to destroy optical transparency and influence the mechanical behavior. Although considerable information is now available on the macroscopic effects of crazing and their relation with other variables such as birefringence, tensile stress, temperature, etc., the basic nature of crazing from the viewpoint of the molecular structure of the material is still not very well understood. It seems advisable, therefore, not only to investigate and record the various external manifestations of this general phenomenon but to examine critically the fundamental nature of crazing from the microscopic and submicroscopic viewpoint. This is the intent of the present investigation which reports the results of observations made with the light microscope, the electron microscope and the x-ray spectrometer. It is hoped that the results of this investigation will aid in providing a better understanding of the crazing of linear high polymers in general.

DEFINITION OF CRAZING AND GENERAL OBSERVATIONS

By "crazing" of thermoplastic materials is meant the development of a state in which the normal optical transparency of the material is reduced. For such thermoplastics as polystyrene or polymethyl methacrylate, crazing may be described in terms of the development of mechanical cracks visible to the naked eye[1] or as a blushing of the material produced by the scattering and reflection of light.[5] Several different patterns of crazing have been observed. Russell[1] has classified them as stress crazing, solvent crazing and combinations of the two. Maxwell and Rahm[4] have likewise reported observations of crazing produced both by application of stress and by solvent action. Sauer, Marin and Hsiao[5] have shown pictures of stress crazing occurring on the fracture surfaces of tension and bending

* This invited paper was presented at the meeting of the Division of High Polymer Physics of the American Physical Society in New York City, February 4, 1950.

[1] E. W. Russell, *Studies on Polymethyl Methacrylate, Part III—Crazing Effects*, Report No. Chem. 447, Royal Aircraft Establishment, Farnborough Hants, England (August, 1948). Nature, London, No. 4186, 165, 91 (1950).

[2] B. Maxwell and L. F. Rahm, Ind. Eng. Chem. 41, 1988 (1949).

[3] J. R. McLoughlin, *Crazing of Polystyrene Films*, Plastics Laboratory Technical Report 12B, Princeton University, Princeton, New Jersey (December 15, 1948).

[4] B. Maxwell and L. F. Rahm, *Factors Affecting the Crazing of Polystyrene*, Plastics Laboratory Technical Report 14B, Princeton University, Princeton, New Jersey (May 5, 1949).

[5] J. A. Sauer, J. Marin, and C. C. Hsiao, J. App. Phys. 20, 507 (1949).

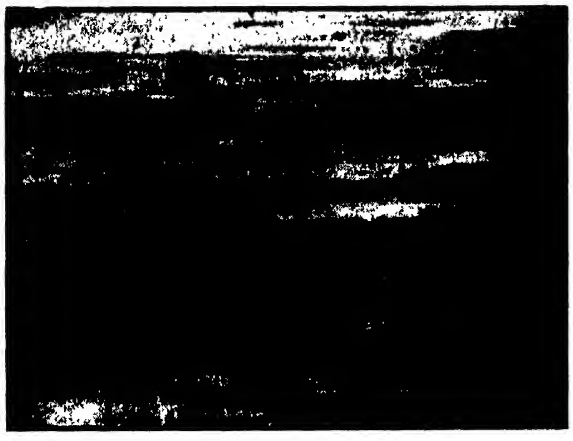


FIG. 1. Crazing cracks in polystyrene.

creep specimens of polystyrene subject to different stresses and different durations of load application. The latter authors have also summarized[5] their observations of the phenomena of stress crazing and interpreted the phenomena observed in terms of changes of molecular structure. It was their feeling that crazing is not merely a mechanical crack phenomenon, although it may be and frequently is closely associated with visible mechanical cracks.

In order to differentiate between these somewhat different descriptions of essentially the same phenomenon, we will consider "crazing cracks" to be the visible cracks produced only after "crazing," in the form of real but submicroscopic discontinuities, has previously come into existence. Crazing may then refer to or be observed in the form of a blushing or haze of a previously transparent material. In other words, crazing is not necessarily actual cracks, but may be some sort of infinitesimal openings which are either still within the field of molecular attraction or are physically prevented by other neighboring molecules from developing into cracks.

The most basic fact about the development of crazing in specimens of linear high polymers subject to load is that the observed crazing cracks are always found to be at right angles to the direction of maximum tensile stress. A microphotograph indicating this behavior for polystyrene is shown in Fig. 1. If, however, the specimen is subject to pure compression rather than tension, no crazing will occur.

Both of the above facts are simultaneously observable when a rectangular specimen is subject to pure bending, as under this type of stress condition one-half of the specimen is subject to tension and the other half to compression. Figure 2 shows two views of a polystyrene specimen which had been continuously subject to a constant bending moment until fracture occurred. The left view clearly shows the absence of crazing on the compression side of the specimen and its strong presence on the tension side. The right view showing the fracture surface reveals on close examination two distinct types of cracks; one type is the fracture cracks which spread radially outward from a more or less point source usually coincident with a flaw, and the second type is the crazing cracks which are seen as normal openings or planes starting from the surfaces of the specimen.

Increase of crazing of polystyrene specimens with increase of time of load application is shown by the six specimens of Fig. 3, each of which was subject to a constant tensile stress varying from 3250 p.s.i. for the specimen on the left to 5500 p.s.i. for the specimen on the right as indicated. The specimens were allowed to remain under these stresses until fracture occurred; for the 5000-p.s.i. specimen the duration of test was very short (minutes), while for the 3250-p.s.i. specimen hundreds of hours elapsed before the specimen failed.

The size, density and depth of penetration of crazing openings depends not only on stress and time but also on the particular material under investigation. For example, it is found that for comparable stress and time conditions, the crazing cracks on the surface of polymethyl methacrylate specimens are usually larger and less dense than those in polystyrene specimens, but at the same time, do not extend inward from the surface as deeply. This latter fact may be due to the

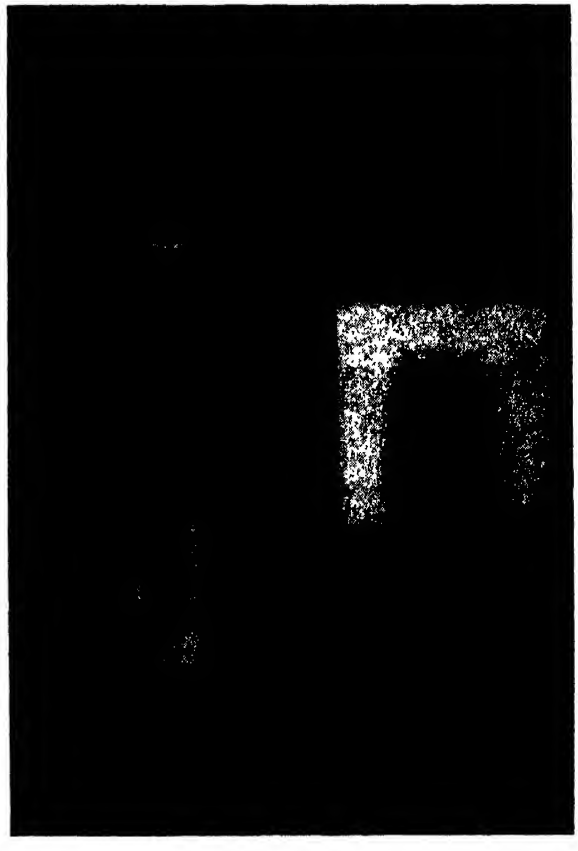


FIG. 2. Bending creep specimen of polystyrene showing crazing along length and on cross section of tension side of specimen.

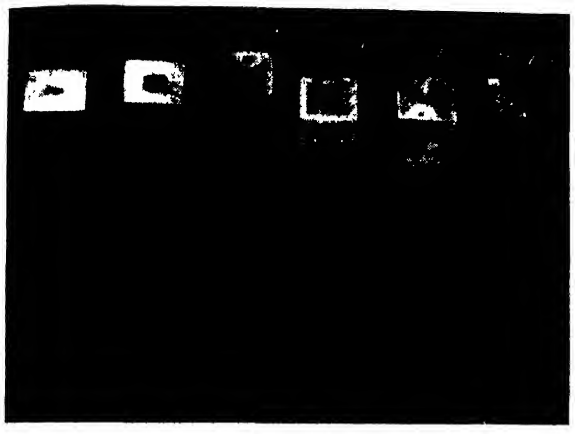


FIG. 3. Fracture surfaces of crazed tension creep specimens of polystyrene subject to different stresses and different durations of load application.

inability of the relatively "brittle" polystyrene specimens to relieve themselves of high local stress concentrations as readily as the more ductile methacrylate specimens.

Another rather interesting observation is that crazing may extend throughout the entire cross-sectional area and yet the specimen can continue to carry load even though reduced in value. This is well shown by the stress-strain curve in Fig. 4 along with that of a standard amorphous polystyrene specimen. While the stress-strain curve of the crazed specimen indicates that it is considerably weaker than the original specimen, it is clear that even the entirely crazed regions of a specimen can carry considerable stress despite the existence of the numerous crazing openings or cracks. The stress-strain curve suggests that the material has been locally oriented due to stretching as the modulus of elasticity of the earlier part of the curve appears to be slightly higher than that of the amorphous material. Above 1500 p.s.i., the crazing openings grow rather rapidly into visible cracks and fracture of the specimen occurs at usual strain values but at lower stress values.

In order to determine whether any similar damage had been done to the apparently non-crazed central portions of a tension creep specimen, the crazed outer portion of the polystyrene specimen was machined off and the remaining central core then subject to tension. The stress-strain curve is shown in Fig. 4 and, on comparison with that of the original amorphous specimen, the change, if any, is seen to be very slight. The modulus of elasticity is slightly higher suggesting that some short-range orientation of the polymer chains might have occurred. Also, a slightly lower ultimate strength might be partly due to the existence of submicroscopic crazing cracks too small to be seen. Both of these effects are, however, small enough to be caused by unavoidable differences between specimens.

In order to find whether compression would be beneficial in eliminating crazing previously produced by tension, a piece of a crazed polystyrene specimen was compressed and released three different times and then tested in tension. Although all the crazing cracks had seemingly disappeared after the third compression, they reopened after tension was applied. The four stress-strain curves are shown in Fig. 5 and the tension stress-strain curve is shown to a larger scale in Fig. 4. While it is clear that the material does not regain completely its original strength, it may be noted that its tensile strength is slightly higher than that of the fully crazed specimen, perhaps indicating that some of the submicroscopic crazing openings have been partially restored thereby providing some additional resistance to loading.

The most noticeable effects of the compressions were the gradual reduction and apparent elimination of the crazing planes and the large increase in ductility.

The effect of the environment on the initiation and development of crazing has been studied by running tensile creep tests on specimens of polystyrene subject

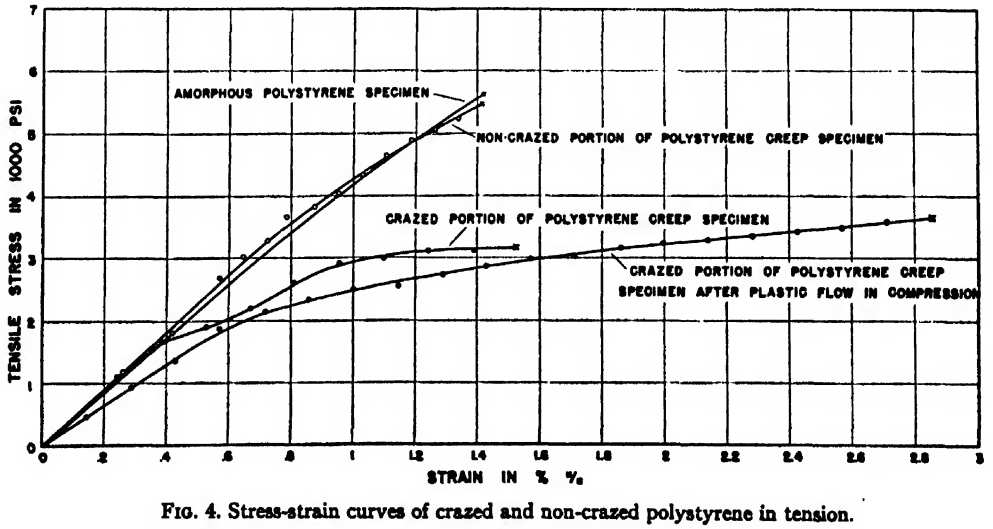


FIG. 4. Stress-strain curves of crazed and non-crazed polystyrene in tension.

to various atmospheric conditions. It was observed that in dry nitrogen, a specimen crazed rather easily and crazing seemed to penetrate deeper than in specimens subjected to the same stress in the ordinary atmosphere (50 percent R.H.). Less crazing was observed for a specimen tested in wet nitrogen. These results seem to indicate that water vapor has a healing effect on cracks that would otherwise open up under the prevalent stress system. This conclusion was subject to additional verification by running one creep test on a polystyrene specimen completely immersed in water. As expected, there was much less penetration of the crazing cracks into the specimen. It was also observed that crazing could be reduced considerably by coating the exterior surfaces of the polystyrene specimen before test. The results were somewhat spotty due probably to the fact that it was rather difficult to get a continuous uniform adherent coating on the square specimen. Nevertheless, the particular coating used, a compound of styrene monomer recommended for attachment of strain gauges to polystyrene specimens, was helpful in reducing crazing.

Crazing can also be produced by action of ultraviolet or x-ray radiation. The effect of the radiation is to cause depolymerization of the material. After the radiation is removed, the material will tend to repolymerize although a fairly long period of time may be required if the temperature is low. On repolymerization, the material will shrink, thereby inducing tensile stresses which in turn will cause crazing cracks to open up at right angles to the stress direction. Since repolymerization is at random, crazing patterns produced by ultraviolet radiation show randomly directed crazing openings or planes and hence are somewhat similar to crazing patterns produced by solvent action.

Although the external manifestations of crazing are now fairly well established for a wide variety of external conditions, little is known about their relation to internal structural changes of the material itself. If crazing could be directly related to elastic or plastic extension of the chain linkages or to tensile or shear deformation between individual chains or to other rearrangements of the polymer chain structure, a major step in general understanding of the phenomena would have been made and would unquestionably lead to fruitful suggestions concerning its elimination.

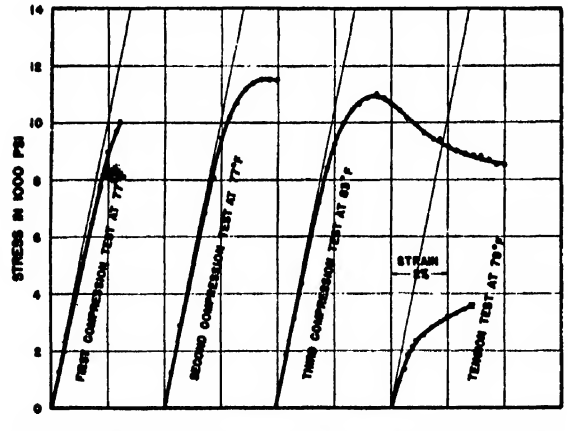


FIG. 5. Effect of compression on the stress-strain curve of crazed polystyrene specimen in tension.

To accomplish the above, the first step would appear to be a careful detailed examination of the crazed and non-crazed surfaces by the best means at our disposal so that whatever changes were present could be detected. This examination is made first visually, then by the light microscope and finally by the electron microscope. In addition, a study has also been made of the crazed and non-crazed portions of the specimen using x-ray diffraction methods. The results of this general examination are now presented.

DETAILED DESCRIPTION OF A PARTIALLY CRAZED POLYSTYRENE SPECIMEN

Lustrex of an average molecular weight between 67,600 and 68,700 was used in this investigation. The stress-strain properties of this material have been described previously by the authors.[5]

The fracture surfaces of tension creep specimens of polystyrene (Fig. 3) subject to different stresses and different durations of load application reveal the dependence of crazing and its growth on the tensile stress magnitude and its duration. The fracture surface is seen to consist of a crazed and a non-crazed portion, the relative amounts of each depending on the magnitude and duration of stress to which each specimen was subjected. The amount of crazing is greater for the specimens subjected to the lower stress levels because the time of load application for these specimens was much greater. The relation between tensile stress and time to fracture appears to be a linear one on a semi-log plot (see Fig. 6).

The specimen at the extreme right of Fig. 3 was loaded at a stress approximately equal to the ultimate strength of polystyrene in tension (~5500 p.s.i.) and hence fractured shortly after application of the load. There is no apparent crazing and no crazing cracks are visible on the fractured surface. The second specimen from the right was stressed at 5000 p.s.i. and hence lasted for some time longer before it failed. The growth of crazing from the edges normally inward toward the center of the specimen is very evident. It may also be noted that the crazing regions are uniformly distributed throughout the total gauge length of the specimen. (This is especially evident when cross-sectional planes other than the fractured surface are examined.) In general, as we go from right to left in Fig. 3, penetration of crazing increases. Meanwhile, the duration of load application also increases in the manner of Fig. 6.

In order to clearly show the difference between crazed and non-crazed material, the fracture surface of a partially crazed specimen of polystyrene was photographed as shown in Fig. 7. This photograph was taken

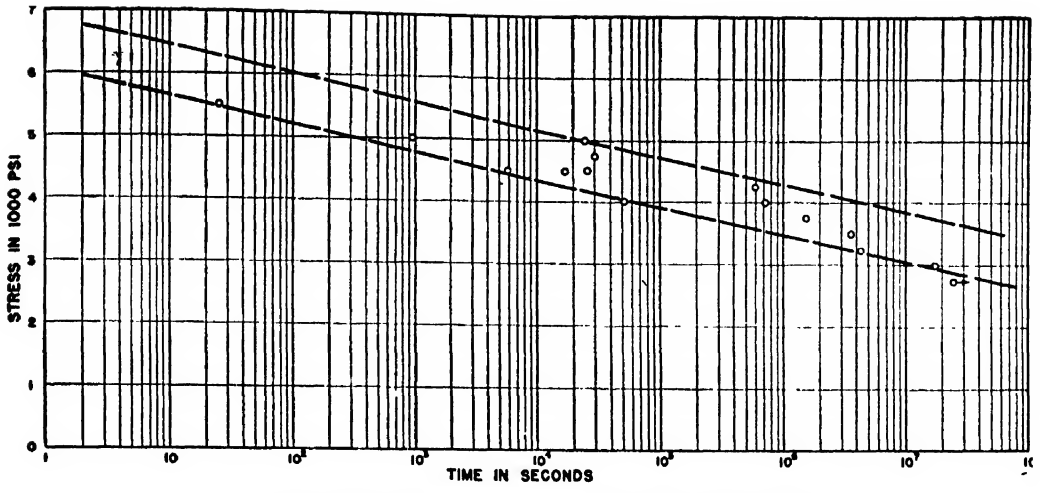


FIG. 6. Effect of tensile stress on fracture time of polystyrene.

by placing the polystyrene specimen against a black background so that the only reflected light would be that coming from the fracture surface. The crazed material along the edges of the specimen is easily differentiated from the centrally located non-crazed material by the smoother, mirror-like appearance.

It should be emphasized that when fracture does occur, the fracture cracks of the material are not the same as crazing cracks and that the source of fracture is usually some flaw in the material and not one of the crazing openings. In Fig. 7, this flaw is located at the center of the smooth circular region. After initiation, the fracture cracks proceed (probably initially at a constant rate) in directions at right angles to the principal stress direction. In Fig. 7, they are seen to extend more or less radially outward from the region in which the fracture first started whereas the crazing cracks are seen to extend normally inward from all edges of the $\frac{1}{2}$ in.$\times\frac{1}{2}$ in. specimen. During the propagation of the crazing and fracture cracks, the effective cross-sectional area of the specimen reduces and hence the tensile stress increases. Accordingly, the molecular rupturing stress is reached at an increasing rate near the neighborhood of the fracture surface. Possibly, the velocity of propagation of the rupture is also increased. Two clear concentric rings just outside the smooth circle indicate that this type of propagation at a varying rate has taken place. Similar phenomena on the fracture of glass have been studied and reported by Smekal.[6]

LIGHT MICROSCOPIC EXAMINATION OF CRAZED POLYSTYRENE

When the fracture surface of a polystyrene tension creep specimen is examined under higher magnifications, there are additional interesting features. Rochow and Rowe[7] have previously reported the characteristic structures of both fractured polystyrene and polymethyl methacrylate surfaces, but did not report on the nature of the crazing regions and surfaces. Under microscopic examination, the blushed or crazed region of Fig. 7 is seen to be entirely different from the amorphous, non-crazed central region. The crazing surfaces are found to be smooth parallel flat surfaces even when examined at a magnification as high as 500×. The flatness of these crazing surfaces as well as their parallel existence throughout the length of the specimens are largely responsible for the high reflectivity of these regions to white light.

Under a magnification of 50×, the fracture surface of a tension creep specimen of polystyrene shows many interesting patterns. Figure 8a is the upper inset of

FIG. 7. Fracture surface of tension creep specimen of polystyrene showing crazed and non-crazed areas and a starting circle from which the cracks are radiated throughout the entire cross-sectional area.

[6] A. Smekal, Glastech. Ber. 13, 141 (1935); Glastech. Ber. 13, 222 (1935); J. Soc. Glass Tech. 20, 432 (1936).
[7] T. G. Rochow and F. G. Rowe, Anal. Chem. 21, 461 (1949).

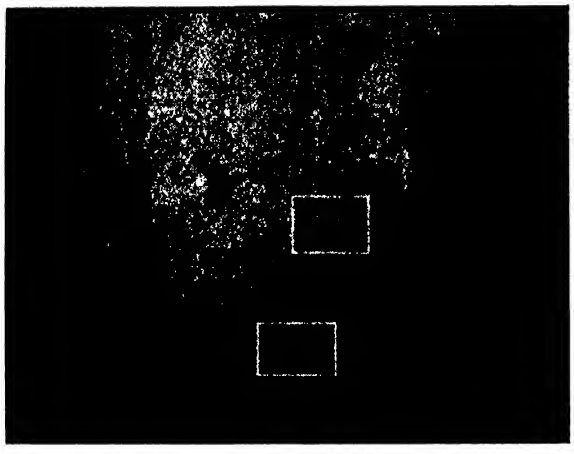


FIG. 8a. Fracture surface of tension creep specimen of polystyrene showing crazed areas and cracks (upper inset of Fig. 7).

FIG. 8b. Fracture surface of tension creep specimen of polystyrene showing the intersection of crazed and non-crazed sections and the starting circle (middle inset of Fig. 7).

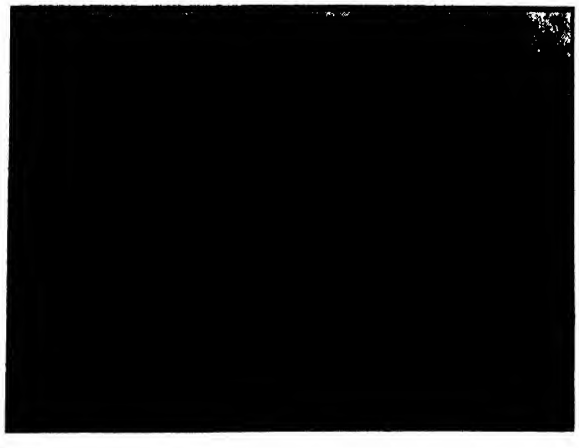


FIG. 8c. Fracture surface of tension creep specimen of polystyrene showing the rings of propagation (lower inset of Fig. 7).

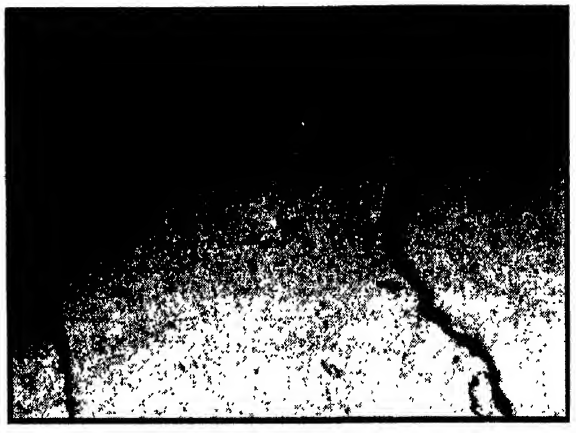


FIG. 9a. Fracture surface of tension creep specimen of polystyrene showing crazed areas and inner structure (upper inset of Fig. 8a).

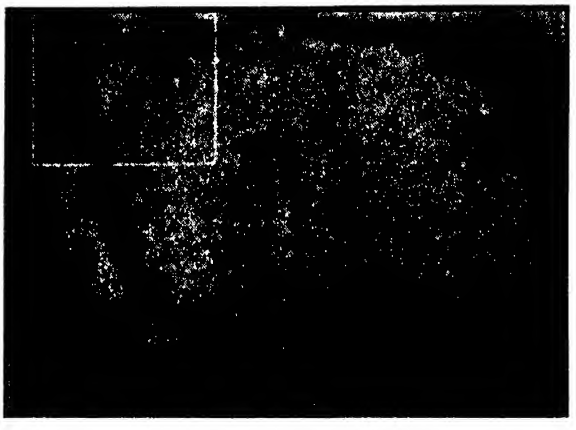


FIG. 9b. Fracture surface of tension creep specimen of polystyrene showing different elevations of crazed areas (lower inset of Fig. 8a).

Fig. 7 which illustrates principally many regions of crazing cracks. These surfaces are quite smooth and reflect white light, if not completely, at least, to a large degree. Both Russell[1] and Maxwell and Rahm[2,4] have made investigations of crazed high polymers with reflected light as the means of determination of the amount of crazing or crazing cracks. It is evident that practically the entire crazed section of a crazed polystyrene specimen can reflect light. The black lines in Fig. 8a are the boundaries of the crazing planes. These planes do not necessarily lie at one elevation as will be seen under higher magnifications later.

The termination of the crazed regions and their juncture with the non-crazed central portion of the specimen is shown in Fig. 8b. The upper white regions are the flat crazing areas separated from one another only by fracture or crazing cracks. The central rupture surface is no longer smooth. As a matter of fact, it is so much different in its elevations that a micrograph at 500× magnification will not show the details of even a very small portion of the texture. Figure 8b shows also,

FIG. 10a. Fracture surface of tension creep specimen of polystyrene showing the connection of edge of the starting circle and the first ring of propagation (left inset of Fig. 8b).

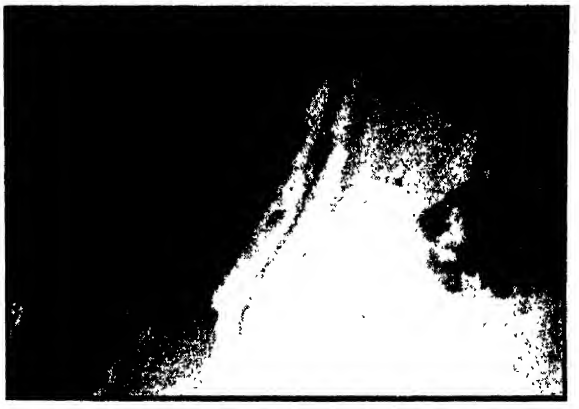


FIG. 10b. Fracture surface of tension creep specimen of polystyrene showing the wavy colored bands occurring inside the starting circle (middle inset of Fig. 8b).

on the right half, part of the smooth starting circle which surrounds the flaw from which the fracture crack propagated. To the naked eye, this circular area is somewhat similar to the crazing area. However, it possesses an entirely different texture. It is a region having needle-like structure with no preferred orientation. Also, to be seen near this region are wavy and continuous lines or contours, indicating the presence of cavities. The clear arc in the middle of this figure is a part of the boundary of the starting circle. Successive rings of propagation are clearly shown in Fig. 8c. As was mentioned before, this is probably indicative of a changing rate of propagation of the crack that continues until failure of the total cross section of the specimen takes place.

The previous pictures have indicated that the surface of the crazing crack is different from that of the rupture crack originating from a flaw. This difference is even more strikingly illustrated by the higher magnification pictures of Figs. 9 and 10. Figures 9a and 9b two are pictures of the surfaces of the crazing areas. In these crazing planes, rather fine structure of the crazing surfaces shows that the smooth mirror-like surface is not a flat plane. Figure 10a shows the fracture surfaces in the region mostly bounded by the first ring of propagation. Figure 10b shows the enlarged cavity and the fine structure in the bottom of the cavity. In Fig. 10c the needle-like structure of the inside region of the starting circle is shown. This structure is somewhat similar to crazing patterns that are induced either by absorption of solvent vapor or by radiation depolymerization. Such patterns likewise show random orientation of the crazing openings. It should be noted that these cylindrical needles or openings are distributed at different elevations, whereas the fine structures on the surface of stress crazing cracks such as those shown in Figs. 9a and 9b are essentially located on one single plane. This latter fine structure is more clearly revealed by still higher magnifications such as Fig. 11 showing the inset of Fig. 9b.

ELECTRON MICROSCOPE EXAMINATION OF CRAZED POLYSTYRENE

In order to show the fine structure of crazed material still more clearly, electron micrographs were prepared. The results of both silica shadowed and beryllium shadowed crazed surfaces of polystyrene show similar structure. Here, only the electron micrograph of a beryllium shadowed crazed fracture surface is illustrated as shown in Fig. 12. This picture apparently correlates very well with Fig. 11. The double lines probably indicate the difference in elevations between two crazing planes. The order of the elevation in this particular picture is about two-tenths of a micron. The smoothness of the crazing surface is estimated to be less than one-tenth of a micron.

It is difficult to obtain a definite interpretation of this picture in terms of the polystyrene molecular structure. Are the small circles, for example, the end views of a group of parallel polymer chains? Under the same conditions of preparation, occasionally figures of the

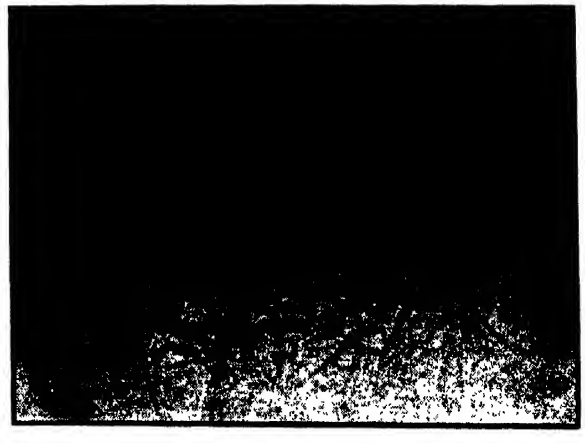


FIG. 10c. Fracture surface of tension creep specimen of polystyrene showing the needle-like structure occurring inside the starting circle (right inset of Fig. 8b).

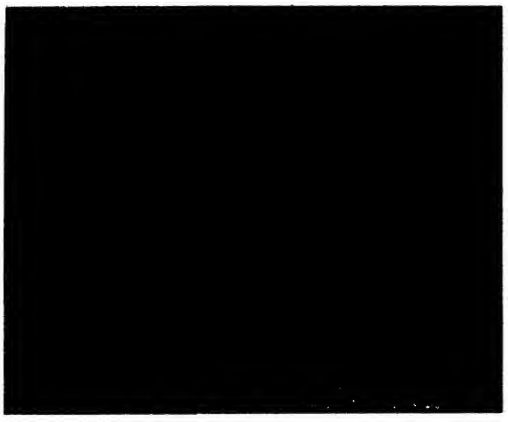


FIG. 11. Fracture surface of tension creep specimens of polystyrene showing the fine structure of crazed area at different elevations (inset of Fig. 9b).

type of Fig. 13 are obtained. Are these pictures of the inverse surface of Fig. 12 or do they indicate the existence of local regions or domains bounded by networks? In contrast to Fig. 12, showing a crazed surface of polystyrene, a molded surface of the same material was also recorded as shown in Fig. 14. This molded surface is perpendicular to the crazing surface. Many fine scratches are shown and the central breakage of the beryllium film may possibly indicate one of the actual large crazing cracks.

In addition to the electron-microscope pictures of the molded and fractured surfaces of solid polystyrene specimens, several additional electron-microscope pictures were taken of oriented polystyrene film. This was done both because the crazing properties of polystyrene are believed to be closely associated with orientation and also to provide some knowledge of the physical behavior of long-chain molecules under their natural configurations. The film was obtained by first dissolving the resin in a solvent and then allowing a drop of the solution to fall into water. As the drop comes back

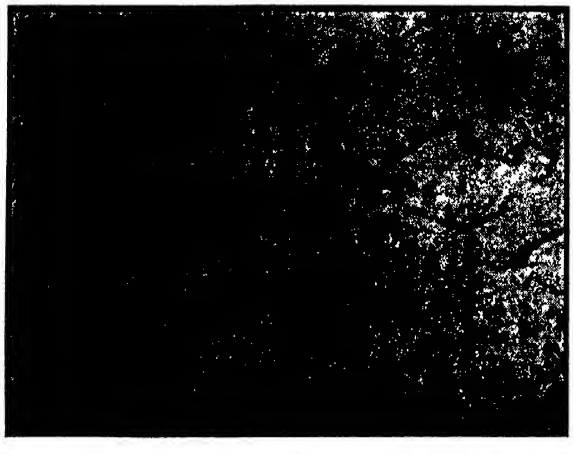


FIG. 12. Electron micrograph of beryllium shadowed crazed polystyrene showing crazing surface at various planes.

FIG. 13. Electron micrograph of beryllium shadowed crazed polystyrene showing crazing surface with various regions of domain structure.

to the water surface, it spreads out radially, the solvent evaporates and the plastic film is left behind on the surface of the water.

Figure 15 is an interesting picture of polystyrene film obtained by dropping polystyrene ethyl bromide solution on a cold water surface. This picture does seem to show the possible existence of domains bounded by networks and hence this type of structure may be the proper interpretation of Fig. 13. Since the formation of domains of polystyrene molecules is probably a very temperature dependent phenomenon, somewhat more striking results might be expected if the ethyl bromide polystyrene solution is dropped in hot water rather than cold. Such indeed is the case. Fig. 16 shows the electron micrograph of polystyrene film when prepared in this manner. This interesting picture reveals the

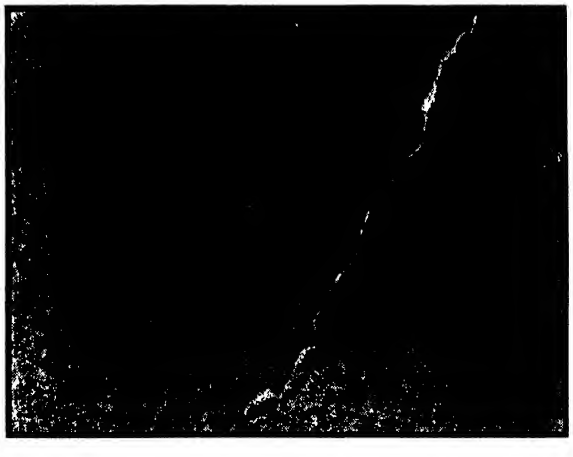


FIG. 14. Electron micrograph of beryllium shadowed polystyrene surface (perpendicular to crazed surface) showing various scratches on the molded surface.

tendency of the material to agglomerate in the form of rods or long slender domains. The picture shows two layers of rodlets both of which are oriented in one general direction. The proper interpretation of this picture is not yet well established but the cylindrical rods are quite possibly more or less parallel pencils of long chain molecules.

The technique employed in the preparation of an oriented film of polystyrene molecules under their natural configurations is briefly described as follows. A drop of dilute polystyrene ethyl bromide solution was dropped in hot water from a position a few inches above the water level. The solution droplet went into the hot water and was heated up before it came back to the water surface. As soon as the solution droplet came out of the water surface, it spread out radially and the polystyrene molecules formed a thin film on the water surface while the ethyl bromide evaporated away. The concentration of the polystyrene ethyl bromide solution varies depending upon the rate of evaporation of this solution. An approximate figure is a few percent of polystyrene by volume in ethyl bromide. The temperature of the hot water is not critical either. For polystyrene, around 80 degrees C, which is the second-order transition temperature of polystyrene, is desirable.

Under this kind of technical treatment, both uniform film and film as shown in Fig. 16 can be obtained. On the basis of this figure and the previous discussion, several generalizations concerning structure may now be attempted:

(a) Under certain conditions, polystyrene molecules can be grouped together as slender domains, (b) It is possible to orient these domains in one general direction, (c) Macroscopic fine structure of polystyrene domains may be observed.

However, there are still many interesting phases which are not understood and need further study. For instance, what is the relationship between the volumetric energy and the surface energy of each domain? Would these domains of polystyrene be present if mercury or other substance was used instead of water, as water molecules are of strong polar groups?

Now, let us consider the size of these rodlets or domains and how they compare with the length of the polymer chains. Figure 16 shows a section of the film magnified some 19,000 times so that a distance of 2 cm would correspond to about 10^{-4} cm or 1 micron. Most of the rods appear to be of the order of 1 micron in length and perhaps one-tenth of a micron in thickness. Now, in polystyrene the primary bonds are C—C bonds having a distance l between them of 1.54A and an angle α between the valence bonds of 109°. The maximum length of an extended chain is then given by[8]

$$L_{\max} = Nnl \sin(\alpha/2),$$

where N is the number of monomer units and n is the number of bonds per unit. For polystyrene, the extended

[8] E. Guth and H. M. James, Ind. Eng. Chem. 33, 624 (1941).

FIG. 15. Electron micrograph of polystyrene film formed by dropping polystyrene ethyl bromide solution on top of cold water surface showing large domain structure.

length of the polymer chain is of the order of 1.6×10^{3}A, or roughly $\frac{1}{6}$ of a micron. Thus, the length of the rodlets is approximately four to six times the length of the polymer chains. The width of the domains is, on the other hand, some 100 times larger than the approximate distance across adjacent chains (10A) and hence if the domain is interpreted as a large conglomeration of more or less parallel chains, it should contain several thousand or more depending on how closely they are packed.

In the polymerized solid state rather than in film form it may well be that the individual long-chain molecules of polystyrene are not arranged in this fashion although this type of structure has been suggested and the polymer chains treated as long rods.[9,10] Regardless

FIG. 16. Electron micrograph of polystyrene film formed by dropping polystyrene ethyl bromide solution in hot water showing oriented domain structures of polystyrene.

[9] H. Mark and G. S. Whitby (Editors), "Collected Papers of W. H. Carothers on High Polymeric Substances," *High Polymers* (Interscience Publishers, Inc., New York, 1940), Vol. I.

[10] K. H. Meyer, "Natural and Synthetic High Polymers," *High Polymers* (Interscience Publishers, Inc., New York, 1942), Vol. IV.

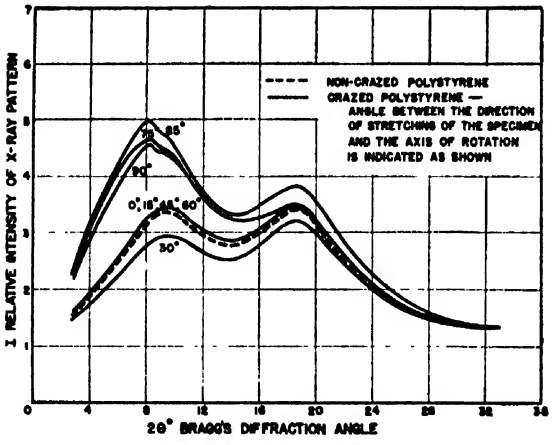


FIG. 17. X-ray photometer curves of crazed polystyrene.

of this situation, Fig. 16 reveals that the natural configuration of polystyrene molecules, at least under certain environments, is one of more or less oriented, intertwined cylindrical domains. Before additional generalizations concerning structure are attempted, results of x-ray investigations of the structure of crazed and non-crazed material will be discussed.

X-RAY INVESTIGATION OF CRAZED AND NON-CRAZED POLYSTYRENE

Both crazed and non-crazed polystyrene samples have been subject to x-ray and electron diffraction examination. The diffraction picture[6] reveals that polystyrene is amorphous in structure. Very little information can be obtained from this picture except to note that the existence of the two somewhat diffuse rings is indicative of the existence of two spacings in some fairly definite regular pattern. Since only one of these rings (the outer one) appears in x-ray pictures of the monomer the other is characteristic of the polymer and may represent either chain to chain spacing or else link to link spacing.

If, however, the x-ray spectrometer is used to obtain intensity measurements, the results indicate a variation of the intensity of the diffracted rings and a shift of the peak of the inner ring or polymerization ring (see Fig. 17). If the peaks indicate some characteristic spacings of the C—C chains, the variation of these spacings with angle is shown in Fig. 18. This shift of the peaks and the variation of the intensity of the diffracted ring for different directions of the axis of rotation relative to the direction of stretching indicate the existence of orientation and "crystallization" of the material.[5,11,12] It should also be noted that no such indications of orientation are found if the non-crazed portions of a specimen are examined by means of the x-ray diffraction.[5]

[11] P. H. Hermans and A. Weidinger, J. Polymer Sci. IV, 135 (1949).
[12] P. H. Hermans and A. Weidinger, J. Polymer Sci. IV, 709 (1949).

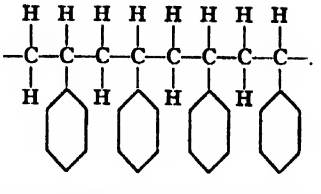
CRAZING AND MOLECULAR STRUCTURE

In the preceding sections, the results of light microscope, electron microscope and x-ray examinations of crazed and non-crazed polystyrene material, as well as of some naturally oriented polymeric film, have been presented. The studies were made on polystyrene or polymethyl methacrylate specimens which had been permitted to craze under the action of stress and time. The crazed material is found to consist of very fine, very smooth more or less parallel openings lying in planes at right angles to the principal tensile stress direction. For a given magnitude and time duration of stressing, the size of the openings and their density varies with the particular polymer under consideration. The openings are finer and much more dense for polystyrene than for methyl methacrylate. Crazing as such is not confined to these two materials, but probably occurs in all linear polymers. It is visible, however, only in transparent materials.

The studies also show that there are differences in orientation between the crazed and non-crazed portions of a linear polymer, although these differences are slight when the crazing is produced by long-time loading at room temperature conditions. There seems also to be some evidence supporting the view previously presented[5] based on observations of behavior of linear polymers under various types of mechanical testing, that polystyrene molecules, at least under certain environments, may group together in more or less cylindrical domain form, with the various long-chain molecules essentially parallel to one another within the domain. The molecular forces across these groups of chains are probably not very large and the precise configuration will depend on which structure gives the minimum free energy. The polymer chains are held together by secondary van der Waals forces and possibly by some side group interlocking.

EFFECT OF DOMAIN STRUCTURE ON CRAZING

Many of the facts pertaining to the inception, growth, and location of crazing and crazing cracks in polystyrene seem to be explainable in terms of orientation and the separation of domains of polystyrene molecules under the action of tensile stress. Consider first the accepted chemical structure of the polystyrene molecule. It is a series of styrene monomers arranged regularly in the following manner



Because of the presence of the large benzene rings around the main C—C chains, it is difficult for these chains to slide in the longitudinal direction of the chains.

However, it should be relatively easy to separate parallel chains by pulling them in their perpendicular directions. It may also be possible under the action of stress for one domain to separate and form several shorter domains if the original domain is many lengths longer than the average length of the long-chain molecules.

The styrene polymer presumably possesses no unused valence bonds except possibly at chain ends. Thus, there are essentially no primary bond linkages across polymer chains and hence the only forces across chains should be relatively weak non-valence or van der Waals London type forces.[13] These forces may be sufficiently large to maintain a rod-like structure when no external forces are acting but under applied transverse tensile stress the polymer chains within the rods would tend to separate from one another. Other evidence that the polymer of styrene is of a thread-like structure arises from consideration of its Raman spectra.[14]

When material having a randomly directed domain type of structure is placed under tensile stress, the various domains tend to orient themselves according to the state of stress applied. Those which are perpendicular to the direction of applied stress will theoretically not be able to orient and in virtue of their weak transverse holding power will open under the action of applied stress. Along the external surfaces of the material these openings will grow more easily than those in the inside, both because of stress concentrations arising from the existence of surface flaws and also because the restrictive force fields of neighboring domains of molecules will probably be somewhat less for domains located on the open surface. When these openings become sufficiently large (many times the dimension of the wave-length of light), they are seen as macrocracks by the naked eye. However, long before the openings become large enough to be individually seen, and when still in the form of microcracks or submicroscopic openings, they may still have the ability to reflect light.[15] In other words, these openings may be sufficiently numerous and sufficiently regular in their general orientation pattern to cause a large proportion of the light striking them to be reflected. Crazing is then evident in the form of surface blushing rather than visible cracks.

When a rigid linear polymer is subjected to tensile stress at room temperature, the tendency of the original randomly directed domains to orient is greatly restricted by the internal molecular interactions of other groups and chains. If, however, the applied stress is high and is maintained for some time, the domains originally

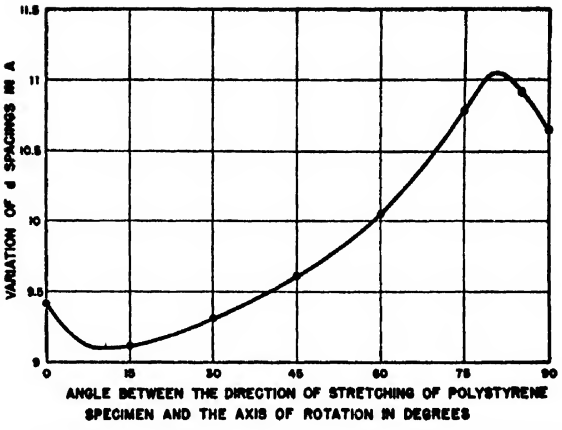


FIG. 18. Variation of polymerization ring of crazed polystyrene.

directed at right angles to the direction of stressing will open up and will initiate the phenomenon of crazing. It is usually possible for many crazing cracks to occur before fracture because the micro-openings soon run up against other domains or groups of molecules not so favorably oriented for separation and then they are automatically halted. A somewhat similar view of the development and growth of crazing cracks has been presented by Maxwell and Rahm.[4]

The mechanism of crazing is now becoming more apparent. Domains or groups of molecules which are at right angles to the applied stress open up first but do not immediately produce failure or a fracture crack as we might expect from prior knowledge of the behavior of brittle materials because they soon run into domains of regions not so oriented. On the submicroscopic or even microscopic scale, the material is by no means homogeneous and the crazing planes are stopped by irregularities in the orientation pattern very much like slip lines in metals are stopped by crystal boundaries lying between differently oriented crystallites. The flatness of the crazing surface (see Fig. 9a) suggests a slow propagation of crazing caused mainly by the joining together of the many microscopic openings induced by separation of the properly oriented domains of the material in their longitudinal directions. The difference in elevation of crazing planes, noted to be about one-fifth of a micron in Fig. 12, is of the same order of magnitude as the extended length of a polymer chain and serves as an indication of the distance the microcrack can extend without running into differently oriented domains.

EFFECT OF ORIENTATION ON CRAZING AND STRENGTH

The effects of orientation on crazing and strength are intimately tied in with the effects of molecular structure. When groups of molecules tend to join together to form domains, it is the opening up of those groups oriented more or less at right angles to the direction of stressing that initiates the development of microcracks or crazing.

[13] A. E. Van Arkel, *Molecules and Crystals* (translated by J. C. Swallow) (Butterworths Scientific Publications, London, and Interscience Publishers, Inc., New York, 1949), Chapter IX.
[14] H. Mark and R. Raff, "High Polymer Reactions, Their Theory and Practice," *High Polymers* (Interscience Publishers, Inc., New York, 1941), Vol. III, p. 73.
[15] S. W. Hawkins and R. B. Richards, J. Polymer Sci. IV, No. 4, 515 (1949).

Also, since the domains can carry considerable stress in their longitudinal direction but only very little stress in the transverse direction, it is clear that the orientation of the domains will greatly affect the strength properties of the polymer.

In general, two kinds of orientation concern us, differing greatly in degree. One is orientation produced by hot stretching and the other by cold stretching. Hot stretching is usually carried out above the second-order transition temperature and only a very small force is required to stretch the material to thousandths of percent of its original length. Evidently, above the second-order transition temperature, thermal agitation energy is great enough to overcome the secondary attractive forces which hold the various polymer chains in more or less tight formation. These chains are then free to slide past one another as well as to orient in the direction of the applied tensile stress. It is thus to be expected that exceedingly large percentages of orientation (in terms of original length) can be produced when the stretching is done at these high temperatures. If the previous picture of the development of crazing is correct, it is also to be expected that when tensile stress is applied in the direction of orientation to specimens previously oriented to some great degree (~1000 percent), crazing, as such, should be absent. This is so because the specimen will no longer contain domains oriented at right angles to the tensile direction and hence no separation of domains is to be expected. Fracture, when it occurs, should be of the fiber type.**

For cold stretching at ordinary room temperature, only little orientation occurs prior to fracture. In polystyrene specimens, failure occurs at a strain of approximately 1.45 percent in./in. and hence only local short range orientation can be accomplished. Somewhat greater orientation might be expected from relatively slow creep tests but even here in terms of percent increase over original length about 1.5 percent seems to be the maximum obtained.

Evidently, the restrictive force fields of neighboring molecules are too great at room temperatures to allow much orientation. That which does occur is also not homogeneous over the entire cross-sectional area of the specimen. On the outside regions of the specimen where surface flaws and the growth of microcrazing cracks affect the stress distribution, x-ray evidence indicates that more orientation takes place than in the central non-crazed portions of the specimen.

The effect of orientation of groups of molecules on strength is a problem which should be susceptible to statistical treatment. It might also be used to check the concept that the polymer molecules are usually grouped together in the form of cylindrical domains bonded together by long-range intermolecular forces with the

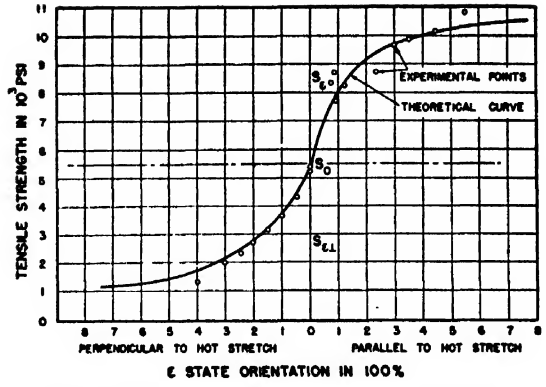


FIG. 19. Effect of uniaxial orientation on ultimate strength of polystyrene.

long chain molecules more or less parallel in each individual domain. For an amorphous polystyrene specimen, these domains of the polymer molecules are randomly oriented. The ultimate strength of this material tested at 77°F±2°F, 50 percent R.H. with the rate of straining 20×10^{-6} sec.$^{-1}$, has been determined experimentally to be approximately 5400 p.s.i. The ultimate strength of oriented material can be shown mathematically to be[16]

$$S_\epsilon = 2\frac{(1+\epsilon)^{\frac{1}{2}}}{1+(1+\epsilon)^{\frac{1}{2}}}S_0, \tag{1}$$

where S_0 is the strength of this material without any preferred orientation, ϵ is the measure of the unit strain in the direction of orientation, and S_ϵ is the ultimate strength of this material oriented at a unit strain change of ϵ in the direction of orientation. When $\epsilon\to0$, S_ϵ is seen to approach S_0 as $(1+\epsilon)^{\frac{1}{2}}/[1+(1+\epsilon)^{\frac{1}{2}}]$ approaches $\frac{1}{2}$. It can also be shown that for the material with perfect orientation in one direction, i.e. $\epsilon=\infty$, S_ϵ approaches $2S_0$ since the expression

$$\left[\frac{(1+\epsilon)^{\frac{1}{2}}}{1+(1+\epsilon)^{\frac{1}{2}}}\right]_{\epsilon\to\infty} = 1.$$

This condition as well as Eq. (1) in general seems to be in satisfactory agreement with experimental studies of oriented polystyrene made by Bailey[17] and hence seems to lend further support to the domain concept of molecular structure.

If oriented material is tested in tension in the direction perpendicular to the direction of orientation, the statistical treatment is somewhat more complicated. The expression found for the ultimate strength is[16]

$$S_{\epsilon\perp} = 2\frac{(1+\epsilon)^{\frac{1}{2}}}{(1+\epsilon)^{\frac{1}{2}}-1}\cdot\frac{2}{\pi}[K-E]S_0, \tag{2}$$

** These observations are in accord with the work of Bailey (see reference 17). Dr. Dietz informs us that similar observations have been experimentally noted in tests on highly oriented polystyrene specimens made at the M.I.T. Plastics Laboratory.

[16] C. C. Hsiao, "Effect of Uniaxial Orientation on Strength and Crazing of Some Linear High Polymers." (It is hoped to submit this paper for publication elsewhere in the near future.)

[17] J. Bailey, India Rubber World 118, 225 (1948).

where K and E are the values of Legendre's complete integrals of the first and second kind respectively. The value $S_{\epsilon\perp}$ varies from S_0 to 0 as ϵ varies from 0 to ∞.

Both Eqs. (1) and (2) for the strength parallel and perpendicular to the stretching direction are plotted in Fig. 19 for various values of percent stretch. It is seen that the predicted results and the experimental results previously reported by Bailey[17] are in good agreement. Small discrepancies that exist between the theoretical and experimental results may be due to differences in rates of straining applied to the specimen.[18]

The general principle employed in the development of the mathematical theory of the effect of orientation on the strength properties of polystyrene is that a domain can contribute its full strength in essentially only one direction, i.e., the direction of its length. Thus, the strength of a linear high polymer is not uniformly distributed from the microscopic point of view. The strength of the material depends upon the orientation of the domains. When stress is applied, some of the unfavorable oriented domains will open up while the more favorably oriented domains will carry the load. The micro-openings gradually grow and join together particularly if the stress is applied for a long period of time. For reasons mentioned previously, they usually start first on the surface and then propagate from the surface of the material toward the center. When they become sufficiently large and numerous enough to noticeably reflect visible light, crazing is said to have occurred.

CONCLUSIONS

A considerable amount of evidence is now available to draw the general conclusion that crazing is an actual mechanical separation of polymer chains or groups of chains under the action of tensile stress. The stress system need not be external as crazing can be produced by internal stresses caused by shrinkage, absorption, radiation, or combinations thereof. Even when the proper stress conditions are present, however, the separations may be too small to be seen either by the naked eye or by a high powered light microscope. The phenomenon of crazing is probably common to many solids other than transparent linear high polymers but, if the material is not transparent, crazing is not observed.

Crazing usually starts on the surface in a region where the polymer chains are oriented at right angles to the direction of applied tensile stress and then gradually penetrates inward. Crazing will not occur if the material is fully oriented in the direction of stretching and it can be reduced for randomly oriented materials by means of suitable surface coatings. Crazing, or even visible crazing cracks, will not necessarily produce fracture but will reduce the total load the specimen can carry. Compression tends to close existing crazing openings and to partially heal them but cannot restore the material to its former state.

For polystyrene, many of the facts pertaining to crazing are explicable in terms of orientation of molecular domains containing many polymer chains held together in more or less parallel fashion by secondary forces. The same type of structure seems to provide a satisfactory model for calculating the effect of orientation on the strength properties of the material.

ACKNOWLEDGMENTS

This project is a joint undertaking with the Materials Branch of the ONR and has been carried on under their sponsorship. The authors wish to acknowledge with many thanks the advice, counsel, and assistance which they received from Dr. W. F. Bartoe of Rohm and Haas Company, and from Dr. T. F. Bates, Mrs. M. V. Black, Dr. K. E. Russell, Dr. W. J. Harrington, Miss E. J. Richeson, and Miss C. Y. Yuan of The Pennsylvania State College.

[18] C. C. Hsiao and J. A. Sauer, "Effect of Strain Rate on the Tensile and Compressive Stress-Strain Properties of Polystyrene," Technical Report No. III, Contract No. N6onr-269, Task Order VI, for Mechanics and Materials Branch, ONR (January, 1950).

End-Cooling of Power Tube Filaments*

JOHN W. CLARK** AND RALPH E. NEUBERT†
Research Division, Collins Radio Company, Cedar Rapids, Iowa
(Received February 24, 1950)

The differential equation defining the relation between temperature and distance in that portion of a vacuum tube filament which is cooled by both conduction and radiation is formulated. This equation has been solved by numerical integration; the results are presented as a set of universal curves in a form convenient for use in vacuum tube design. The calculations agree well with some experimental measurements which were made to verify the theory.

INTRODUCTION

IN conventional power tube design the end cooling of the filaments by their supports is not particularly important and is usually treated by simple approximate methods.[1] This is due to the fact that the filaments in such tubes are very long compared to their diameter, so the power lost by conduction is much smaller than that radiated. Power tubes intended for use at very high frequency must have their filaments not too long as compared with the shortest wave-length at which they will be operated; accordingly as the radio industry continues to demand more power at high frequency the tube designer is forced to use correspondingly short filaments in which conduction cooling cannot be neglected.

This paper presents a method for designing the end portion of a filament so that the temperature distribution in that portion of the filament affected by conduction can be readily determined. The solutions are presented as a family of universal curves for convenience in design work. While this method is intended primarily for use in connection with the design of short filaments, it may be used for filaments of any length.

FORMULATION OF THE PROBLEM

In order to fix the ideas, let us first consider the simple case of a filament of cross section area A, perimeter P, and resistivity ρ which is cooled by radiation only. Then in any incremental length of filament the power lost by radiation must equal the Joule heat developed in that same increment of length by the current passing through it. That is

$$I^2R=I^2(\rho dx/A)=P\gamma dx, \tag{1}$$

where I is the current passing through the filament and γ is the radiation coefficient, i.e. the number of watts per square centimeter radiated by the filament. The equation is set up in terms of P, the perimeter, so it will apply regardless of the shape of the wire. It is convenient to write this equation in terms of the current density $J\equiv I/A$, and of a shape factor $F\equiv P/A$. In terms of these new variables,

$$J^2\rho=F\gamma. \tag{2}$$

This simple equation determines the operating temperature of a radiation-limited filament. Ordinarily it is required to determine the current to produce a given temperature in a wire of given cross section. Since both ρ and γ are known as functions of temperature for any material which might be used to construct a filament, it is a straightforward matter to solve the problem.

Let us now consider that portion of the wire which is so close to the support that it is not at uniform temperature. This portion of the wire will lose heat both by radiation and by conduction. Figure 1 defines the geometry of the problem. The heat carried across a boundary at x must be the difference between the Joule heat generated and the heat lost by radiation between x and 0. The point $x=0$ is defined as the value of x where $dT/dx=0$. Beyond this point, radiation alone suffices to carry away the I^2R heat, as formulated in Eq. (2). This energy balance can be formulated in the following equation, letting K be the coefficient of thermal conductivity:

$$\int_0^x \frac{I^2\rho(T)}{A}dx-\int_0^x P\gamma(T)dx=AK(T)\frac{dT}{dx}. \tag{3}$$

Rewriting (3) in terms of current density and shape factor, we have the fundamental integro-differential equation of temperature distribution in a filament cooled by both conduction and radiation:

$$\int_0^x J^2\rho(T)dx=\int_0^x F\gamma(T)dx+K(T)\frac{dT}{dx}. \tag{4}$$

Now $\rho(T)$, $\gamma(T)$, and $K(T)$ are all known functions of temperature. Accordingly we can differentiate (4) to get it in the form of a second order differential equation:

$$J^2\rho(T)=F\gamma(T)+K(T)\frac{d^2T}{dx^2}+\frac{dK(T)}{dx}\frac{dT}{dx}. \tag{5}$$

The remainder of this paper is concerned with obtaining solutions of Eq. (5).

* This work was supported in part by the United States Air Force through a subcontract between Collins Radio Company and the RAND Corporation.
** Now at Varian Associates, San Carlos, California.
† Now at Sylvania Electric Products, Inc., Emporium, Pennsylvania.
[1] L. R. Koller, *The Physics of Electron Tubes* (McGraw-Hill, Book Company, Inc., New York, 1937), pp. 89–91.

SOLUTION OF THE TEMPERATURE EQUATION

The formal solution of Eq. (5) is

$$x(T)=\int_T^{T_H}\frac{K\,dT}{\left(\int_T^{T_H}2K(J^2\rho-F\gamma)dT\right)^{\frac{1}{2}}}. \qquad (6)$$

The solution of Eq. (5) is outlined in Appendix A.

Since ρ, γ, and K are all known functions of temperature, it is in principle possible to evaluate the integrals in (6) for any case of interest. This is a very laborious process and hence it is useful to calculate a set of generalized curves which can readily be used in any particular case.

The only metal for which detailed data are available for values of electrical resistivity, thermal conductivity, and specific radiation over the temperature range of interest is tungsten. Fortunately this is the metal most likely to be used for a power tube filament.

Equation (6) has been solved by numerical integration using the known values of the physical properties of tungsten. The numerical integration was performed by the Computations Group of the RAND Corporation, Santa Monica, California. This assistance is gratefully acknowledged.

The method used in the numerical integration is given in Appendix B. The results were computed with the aid of an auxiliary function K^* defined by the following equation:

$$x(t)=(10^4/J)K^*(t, t_H, \alpha), \qquad (7)$$

where t is temperature measured in kilo-degrees centigrade and the parameter α is the shape factor F divided by the square of the current density in amperes per square centimeter. That is

$$\alpha\equiv 10^6F/J^2. \qquad (8)$$

These calculations enable one to predict the temperature distribution in a tungsten filament in the range of interest in typical power tubes. Typical curves are shown in Fig. 2 for one specific case. The agreement between calculation and experiment is quite gratifying.

One is frequently interested in the distance from the cold filament support to the point where radiation cooling becomes controlling, rather than in the detailed variation of temperature up to this point. This point has previously been identified as the point where $T=T_H$; it is here that dT/dx vanishes and the useful part of the filament (as far as electron emission is concerned) starts. Let us denote the value of x corresponding to $T=T_H$ by l. Then l is the length of that part of the filament in which conduction cooling causes a variation in temperature.

It frequently happens that l is fixed by other design considerations. Since l is an implicit function of t_H, F, and J, it is possible to use the calculated values to determine values of F and J consistent with the given

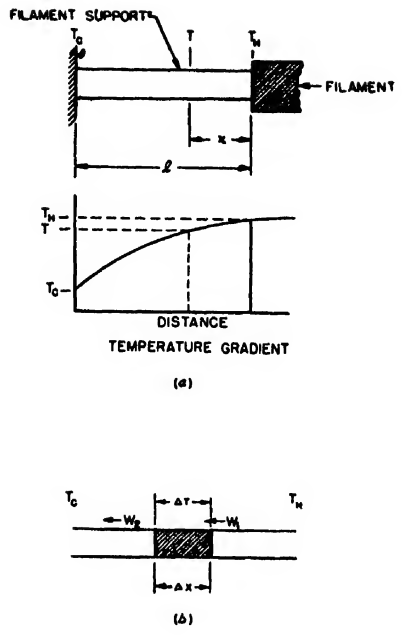


FIG. 1. The geometry of the problem.

values of t_H and l. To facilitate this process, the curves of Fig. 3 were prepared. An inspection of Eq. (7) shows why this single family of curves can give all the required information. Remembering that $l\equiv x(t_H)$, we obtain from (7):

$$l=(10^4/J)K^*(t_H, \alpha). \qquad (9)$$

Thus it is possible to plot one family of curves with lJ as ordinate, with α as abscissa, and with T_H as a parameter. It is convenient to use $(lJ)^{-1}$ as ordinate rather than lJ; this has been done in Fig. 3. On this plot the heavy lines are fixed values of T_H, the temperature of the emitting portion of the filament. The abscissa is $\alpha\equiv 10^6F/J^2=10^6PA/I^2$. The ordinate is $10^4(Jl)^{-1}=10^4A/Il$. The light lines are lines of constant $100Pl/I$.

One can enter the plot with given values of any four of the five quantities T_H, P, A, I and l and obtain the corresponding value of the fifth. Usually other considerations fix T_H, I, and l; and if the shape of the wire is fixed, P and A are of course not independent. This means that it may be necessary to make either the shape or the area of that portion of the filament where temperature variations exist different from that of the filament proper. This can be regarded as a procedure for controlling the position of the beginning of the emitting region; the curves of Fig. 3 give one all the information necessary to do this.

DETERMINATION OF PHYSICAL CONSTANTS

The method outlined above can also be used to determine the three physical constants ρ, K, and γ of metals at elevated temperatures. The writers have not done this, but feel that a brief outline of the procedure would be appropriate. In view of the paucity of data concerning the physical properties of metals at high

temperatures, it would be very desirable for someone to undertake these measurements.

A test piece (of circular cross section) at least ten diameters long would be mounted between electrodes of accurately controlled temperature and of accurately known spacing. Data would be obtained of brightness temperature and current as a function of applied voltage. A new sample of the same wire would then be mounted in the same way, with the electrodes moved so that the sample is now only about four diameters long; it must still have a short radiation-limited area at its center at the highest operating temperature. The same data as before are taken. It is now easy to obtain resistance *vs.* temperature curves for each sample, and by subtraction to obtain resistance *vs.* temperature for the mid-portion of the filament which was cooled by radiation only. Thus we can calculate ρ as a function of temperature, and by the use of Eq. (2) we can calculate the specific radiance γ.

If we are dealing with a metal for which the relation between brightness temperature and blackbody temperature is not known, this will have to be determined in a separate experiment.

The determination of K, the thermal conductivity, can be accomplished with the same set-up. We now use a still shorter wire, so that radiation limiting does not occur. We then measure the mid-temperature, and the distance Δx from the midpoint of the wire to a point where the temperature is ΔT degrees less than the mid-temperature. Writing Eq. (4) as a difference equation, and solving for K, we find

$$K=(J^2\rho-F\gamma)(\Delta x)^2/2\Delta T. \tag{10}$$

The factor 2 in the denominator occurs because we are averaging K over the interval in question.

APPENDIX A

To solve the second-order non-linear differential equation

$$J^2\rho(T)=F\gamma(T)+K(T)\frac{d^2T}{dx^2}+\frac{dK(T)}{dx}\frac{dT}{dx}, \tag{5}$$

where $\rho(T)$, $\gamma(T)$, and $K(T)$ are known tabulated functions.

Rearranging,

$$\frac{d^2T}{dx^2}+\frac{K'(T)}{K(T)}\left(\frac{dT}{dx}\right)^2=\frac{J^2\rho(T)-F\gamma(T)}{K(T)}. \tag{11}$$

We have written K' for dK/dT and have made use of the fact that $dK/dx=dK/dT\cdot dT/dx$. Now let $dT/dx=u$; then $d^2T/dx^2=udu/dt$, and our equation becomes

$$u\frac{du}{dT}+\frac{K'(T)}{K(T)}u^2=\frac{J^2\rho(T)-F\gamma(T)}{K(T)}. \tag{12}$$

Let $m=u^2$; then $udu/dt=\frac{1}{2}dm/dt$, and

$$\frac{dm}{dt}+2\frac{K'(T)}{K(T)}m=2\frac{J^2\rho(T)-F\gamma(T)}{K(T)}. \tag{13}$$

This is a linear differential equation of the first order. The solution is:[2]

$$m[K(T)]^2=2\int_T^{T_H}k(T)[J^2\rho(T)-F\gamma(T)]dT. \tag{14}$$

Returning to our original variables,

$$\left(\frac{dT}{dx}\right)^2=\frac{2}{[K(T)]^2}\int_T^{T_H}K(T)[J^2\rho(T)-F\gamma(T)]dT. \tag{15}$$

Equation (15) can be formally solved for x, giving the integral

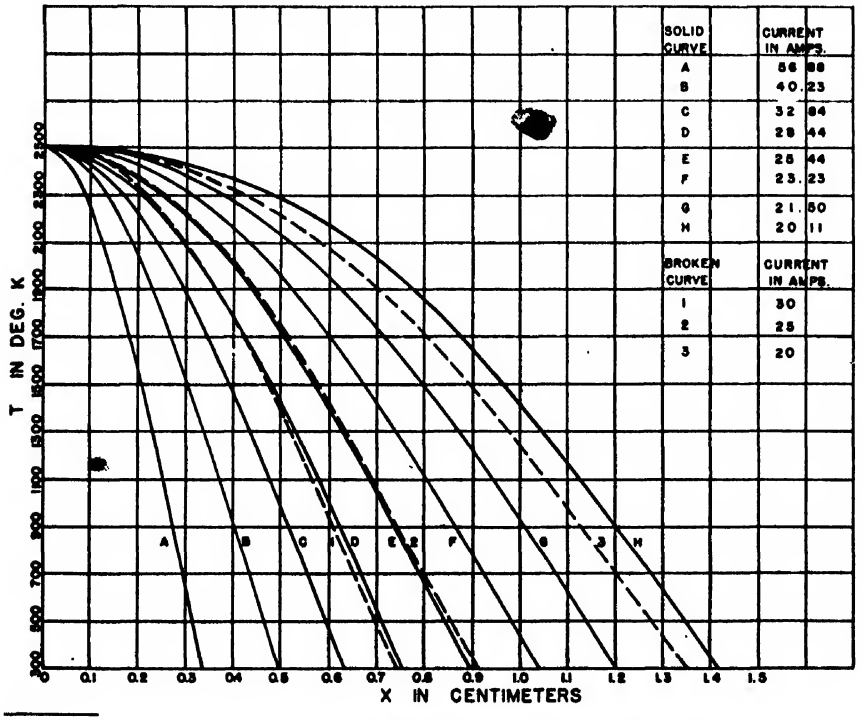


FIG. 2. Comparison between measured and calculated values of temperature distribution in a tungsten filament 0.020 in. diameter. The solid curves are calculated, while the dashed curves are measured values.

[2] F. S. Woods, *Advanced Calculus* (Ginn and Company, Boston, 1932), p. 220.

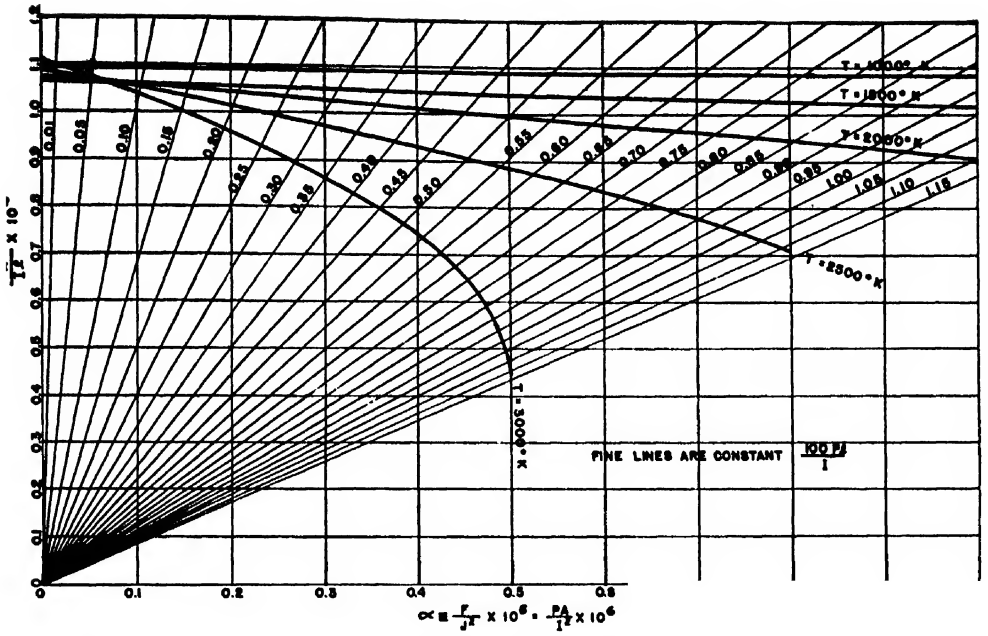


FIG. 3. Universal curves for determining the length of the conduction cooled portion of a tungsten filament.

solution

$$x(T)=\int_T^{T_H}\frac{K(T)dT}{\left(\int_T^{T_H}2K(T)[J^2\rho(T)-F\gamma(T)]dT\right)^{\frac{1}{2}}}. \tag{6}$$

APPENDIX B

The numerical integration of Eq. (6) was carried out as follows:

The physical constants of tungsten as functions of temperature can be approximated by the following equations:

$$K(T)=1.346-1.6\times10^{-4}T, \tag{16}$$

$$\gamma(T)=1.818\times10^{-19}(\log T)^{38.8}, \tag{17}$$

$$\rho(T)=1.51\times10^{-6}\left(\frac{T}{100}\right)^{1.21}. \tag{18}$$

The data from which these empirical equations were derived was obtained from Jones and Langmuir[3] for electrical conductivity and radiation constants, and from Coolidge[4] and Osborn[5] for thermal conductivity.

The analytic functions are introduced into the integrand of (6) and the integration performed by machine. In order to present the results compactly, two new variables are introduced:

$$t=T/1000, \tag{19}$$

$$\alpha=10^6F/J^2. \tag{20}$$

It is apparent from the form of (6) that it can be written

$$x(t)=(10^4/J)K^*(t,t_H,\alpha); \tag{7}$$

Eqs. (6) and (7) together constitute the definition of the function K^*. This function was computed with the aid of the approximate analytical formulations given in (16), (17) and (18) for $T_H=1000°K$, $1500°K$, $2000°K$, $2500°K$, and $3000°K$, for appropriate fixed values of α. The results were used in plotting the design curves given in Fig. 3.

---

[3] H. A. Jones and I. Langmuir, "The characteristics of tungsten filaments as functions of temperature," Gen. Elec. Rev. 30, 312–313 (June, 1927).

[4] *Handbook of Chemistry and Physics*, 30th edition (Chemical Rubber Publishing Company, Cleveland, Ohio, 1947), p. 1865. Converted to watts/cm²/°C from Coolidge's data at 18°C.

[5] R. H. Osborn, J. Opt. Soc. Am. 31, 428–431 (June, 1941). There appears to be a genuine discrepancy between the room temperature value of K and that at elevated temperatures. The empirical Eq. (16) fits Osborne's data and does not pass through the point given at room temperature. Further work is needed to resolve this discrepancy.

Space-Charge Effects in Electron Beams and Their Reduction by Positive Ion Trapping*

E. G. Linder and K. G. Hernqvist**
RCA Laboratories Division, Radio Corporation of America, Princeton, New Jersey
(Received April 19, 1950)

The spread of electron beams caused by space-charge forces is discussed. Under certain conditions ions formed from the residual gas may be trapped in the beam. An equilibrium condition may be established in which the electron and ion densities are equal, and then space-charge forces will be neutralized. Under these conditions high current densities may be produced at low voltages. A theory of ion trapping is discussed, and the equilibrium condition is formulated.

Experimental data are presented which were obtained by the application of microsecond pulses to the beam. This technique is described, and its advantages and possibilities are mentioned. Data are given on beam spread as a function of current, voltage and pressure. Data on the improvement due to ion trapping are included. An increase of current density by a factor of 30 was observed with the structures tested. Results are included on ion build-up time as a function of pressure, and on beam noise and stability in the presence of trapped ions.

I. INTRODUCTION

SPACE charge is the chief limiting factor in the production of high current-density electron beams. On account of space charge, repulsive forces exist so that beams parallel at their origin, become divergent, or beams originally convergent fail to converge. This effect appears to have been studied first by E. E. Watson[1] who, in 1927, derived the expression for the divergence of an initially parallel beam. Numerous other workers[2–8] have further developed Watson's results, and some have devised methods of controlling the spread. In 1940 Thompson and Headrick[9] extended the theory to the case of convergent beams. Field, Spangenberg, and Helm[10] showed in 1947 that a universal curve could be drawn giving beam spread in terms of certain general parameters.

This article covers the following topics: (a) Theory of space-charge spread, (b) Ion trapping and space-charge neutralization, (c) Pulse technique and apparatus, (d) Experimental data, (e) Beam stability.

M.k.s. units are used throughout, except when otherwise specified. The abbreviation SCN will be used throughout for "space charge neutralized," or "space charge neutralization."

II. THEORY OF SPACE-CHARGE SPREAD

Thompson and Headrick[9] have derived an expression for the spread of a convergent beam giving the shape, or profile, for systems of circular transverse symmetry. They found that the radius of the beam at any point is

$$r = R_0 \epsilon^{-\pi\epsilon_0 \frac{m}{e} \frac{v}{I}(v_{ro}^2 - v_r^2)},$$
$$= R_0 \epsilon^{-S(V_{ro} - V_r)/V}, \quad (1)$$

where R_0 = initial beam radius, I = beam current, ϵ = natural logarithmic base, ϵ_0 = permittivity of free space, m = electron mass, e = electron charge, v = beam velocity, v_r = radial velocity, v_{ro} = initial radial velocity, V = beam voltage, V_r = radial velocity in equivalent volts, V_{ro} = initial radial velocity in equivalent volts, and $S = (8\pi^2\epsilon_0^2 e/m)^{\frac{1}{2}} V^{\frac{1}{2}}/I$. The minimum radius is obtained by putting $V_r = 0$, that is,

$$r_m = R_0 \epsilon^{-SV_{ro}/V}. \quad (2)$$

The longitudinal (axial) position of any surface electron is found to be

$$z = 2S^{\frac{1}{2}} R_0 \epsilon^{-SV_{ro}/V} \int_{(SV_r/V)^{\frac{1}{2}}}^{(SV_{ro}/V)^{\frac{1}{2}}} \epsilon^{x^2} dx, \quad (3)$$

and the axial position of the point of minimum radius is

$$z_m = 2S^{\frac{1}{2}} R_0 \epsilon^{-SV_{ro}/V} \int_0^{(SV_{ro}/V)^{\frac{1}{2}}} \epsilon^{x^2} dx. \quad (4)$$

A typical set of curves computed from Eqs. (1) and (3) is given in Fig. 1. Profiles b are plotted for a series of beam currents from 0 to 60 ma. The beam voltage is 300, the source radius $R_0 = 0.3$ cm, and the focal distance with no spreading is $z_0 = 0.8$ cm. The curve cc is the locus of the points of minimum radius.

In the same figure, a partition A is shown which contains a circular aperture a, concentric with the beam. By use of the beam profiles it is evident that the amount of current passing through this aperture may be computed. Thus it is seen that for currents up to about 9.0 ma, the total amount traverses the aperture, whereas at 20 ma less than 25 percent passes through, and at 60 ma only about 3.5 percent.

* Presented at the Symposium of the Division of Electron Physics, Cambridge, Mass., June 18, 1949.
** Scandinavian-American trainee.
[1] E. E. Watson, Phil. Mag. 3, 849 (1927).
[2] M. Knoll and E. Ruska, Ann. d. Physik 12, 607 (1932).
[3] F. D. Fowler and G. E. Gibson, Phys. Rev. 46, 1075 (1934).
[4] A. Bouwers, Physica 2, 145 (1935).
[5] B. von Borries and J. Dosse, Arch. f. Electroteck. 32, 221 (1938).
[6] A. V. Haeff, Proc. I.R.E. 27, 586 (1939).
[7] L. P. Smith and P. L. Hartman, J. App. Phys. 11, 220 (1940).
[8] J. R. Pierce, J. App. Phys. 11, 548 (1940).
[9] B. J. Thompson and L. B. Headrick, Proc. I.R.E. 28, 318 (1940).
[10] L. M. Field, K. Spangenberg, and R. Helm, Elec. Communication 24, 108 (1947).

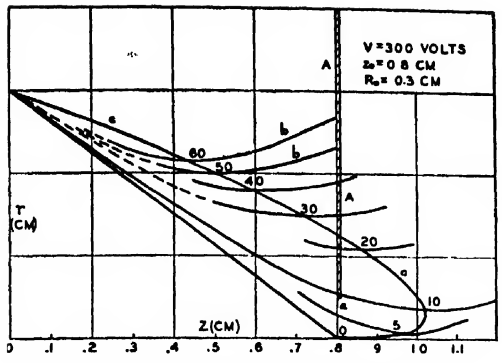


FIG. 1. Typical computed beam profiles.

If I_a is the current traversing the aperture, and I is the total beam current, the beam efficiency may be defined as I_a/I. This is determined by the ratio of the aperture area to the beam cross-sectional area, i.e.,

$$\mathcal{E}=I_a/I=r_a^2/r_1^2,$$

where r_a is the aperture radius and r_1 is the beam radius at the aperture.

A set of curves showing beam efficiency, thus computed, is given in Fig. 2, for beam voltages of 200, 300, and 400 volts. These curves, as well as those of Fig. 1, are idealized in that they are based on the original Thompson and Headrick assumptions, and furthermore no account has been taken of any aberrations, or effects of initial velocities.

By inspection of Eqs. (2) and (4) it is seen that r_m and z_m may be written as functions of the quantity $V^{\frac{1}{2}}/I$, since, by assumption, V_{ro} is proportional to V. It is not evident by inspection that Eqs. (1) and (3) also may be written as functions of this same parameter. However it has been shown by Field, Spangenberg, and Helm[10] that universal beam profiles may be drawn in terms of this parameter. Hence curves such as those of Fig. 1, may be considered universal, in so far as one may be drawn for each value of $V^{\frac{1}{2}}/I$, for any given geometry.

Thus, the curves of Fig. 2, also may be combined into a single universal curve by plotting $I/V^{\frac{1}{2}}$ along the horizontal axis instead of I. Such a curve is shown in

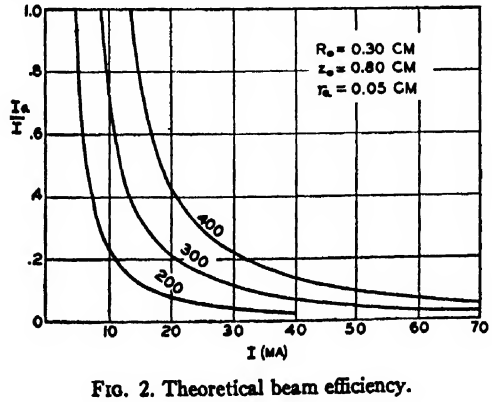


FIG. 2. Theoretical beam efficiency.

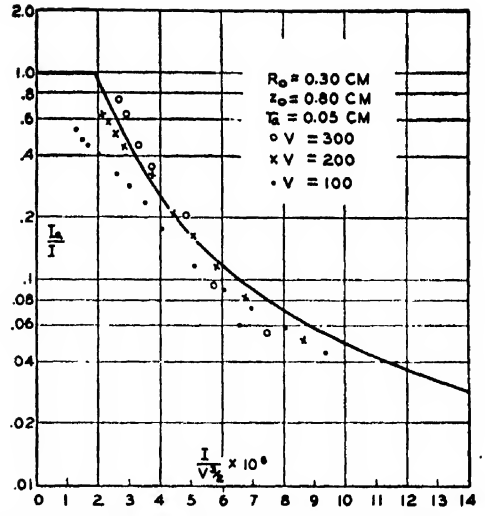


FIG. 3. Universal beam efficiency curve with experimental points at three different voltages.

Fig. 3. The experimental data shown in this figure will be discussed later.

III. ION TRAPPING AND SPACE CHARGE NEUTRALIZATION

In this section the prevention of space charge spread will be discussed. The method used here is based on the neutralization of electron space charge by positive ions. The first similar application of positive ions appears to have been made by J. B. Johnson[11] in 1922 in connection with a cathode-ray oscillograph tube in which beam spread was reduced by admitting gas into the tube. However, the conditions differed widely from those under discussion in this article, and did not involve the trapping of ions as described hereinafter.

A more recent application by O. Heil[12] has been described but many details are not available. The most detailed study was made by Field, Spangenberg, and Helm.[13] Here ions formed by collisions between beam electrons and residual gas molecules were trapped in long, high voltage beams. It was also shown that the number of residual gas molecules at pressures used in ordinary high vacuum tubes was sufficient to supply ions for beams of high current density. However, again, conditions and methods differed considerably from those of the present work.

An understanding of the ion-trapping process may be had with the aid of Fig. 4, which shows the potential variation across the tube diameter. At the top, a cross section of a cylindrical tube and beam is illustrated, where D is the cylindrical wall, and b is the beam, see

[11] J. B. Johnson, J. Opt. Soc. Am. 6, 701 (1922).
[12] O. Heil, *German War Time Research and Development in Klystrons*, U. S. Dept. of Commerce, Publication Board Report No. 52,348.
[13] See reference 10. This same work was printed privately in 1942. See report on "The Production and Control of Electron Beams," Electrical Engineering Dept., Stanford University.

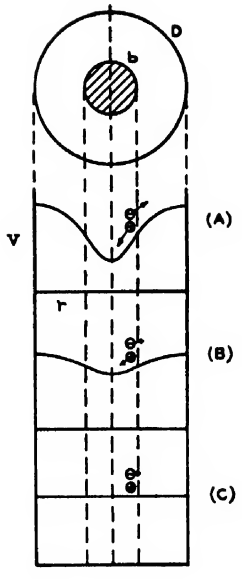


FIG. 4. Potential distribution across beam with various degrees of SCN.

Fig. 6. Below at (A) is shown a typical potential variation for an unneutralized beam. In the case of a cylindrical, parallel beam the variation is approximately parabolic within the beam, and logarithmic between the beam and the wall, (see Smith and Hartman[7]). Thus there is a potential minimum at the beam center.

This depression extends along the longitudinal axis, but decreases as the grid or aperture are approached, and becomes negligibly small at these electrodes. Thus there is formed a potential valley surrounded by higher regions on all sides.

Such a potential valley acts as an ion trap, as shown in Fig. 4(A). Positive ions, formed by collisions between beam electrons and residual gas molecules, move toward the beam center, and are trapped. On the other hand, low velocity electrons, formed by the ionization process, are forced away from the center of the beam.

The resulting accumulation of ions in the beam causes a progressive neutralization of electron charge, and consequently, the potential rises and the depression decreases, as shown at (B). If, eventually, the positive ion density equals the electron density, the potential becomes uniform, as at (C), then ion-trapping obviously ceases, and an equilibrium condition is established. Actually, under most circumstances, there will remain a slight potential depression of the order of a few tenths of a volt just sufficient to overcome the thermal energies of the ions. This will be discussed in detail below. Furthermore, electrons formed by ionization will continue to be lost from the beam, even with no potential depression, since they will encounter no resisting forces.

When the electron space charge is thus neutralized, the electron paths become straight and radial. The total current then passes through the aperture, except insofar as there are limitations due to unconsidered factors, such as structural aberrations and initial velocities.

The rate of formation of ions in an electron beam is given by

$$(dN_2/dt)_1 = \pi r^2 p P(V) n_1 v_1, \tag{5}$$

where p is the gas pressure, P is the number of electrons (or ions) formed per unit length, per electron, at unit pressure, n_1 is the electron density, N_2 is the number of ions per unit beam length, and v_1 is the electron velocity.

The rate of loss of ions will be computed by assuming that the trapped ions form an ion gas. This seems plausible in view of the long lifetime of the ions in the trap and the consequent possibility of the attainment of an equilibrium state. The rate of loss will then depend upon the ion gas density at the wall and the average normal velocity component there. The density will be controlled by the Boltzmann law and at the wall it will be

$$n_2 = n_2' \epsilon^{\Delta V e/kT_2},$$

where n_2' is the ion density at the beam center, T_2 is the gas temperature, and ΔV is the potential difference between the beam center and surrounding cylindrical wall. In the case of an electron beam, the potential is depressed, and ΔV is such as to make $\Delta Ve/kT_2$ negative for positive ions. The average normal velocity at the wall is, from kinetic theory,

$$v_2 = \left(\frac{V_{02}e}{2\pi M}\right)^{\frac{1}{2}}.$$

Thus the rate of ion loss may be written

$$\left(\frac{dN_2}{dt}\right)_2 = 2\pi R n_2' \epsilon^{\Delta V/V_{02}} \left(\frac{V_{02}e}{2\pi M}\right)^{\frac{1}{2}}. \tag{6}$$

The net rate of change of number of ions may therefore be written

$$\frac{dN_2}{dt} = \left(\frac{dN_2}{dt}\right)_1 - \left(\frac{dN_2}{dt}\right)_2 = \pi r^2 p P n_1 v_1 - 2\pi R n_2' \epsilon^{\Delta V/V_{02}} \left(\frac{V_{02}e}{2\pi M}\right)^{\frac{1}{2}}. \tag{7}$$

Throughout the trapping period the first term remains constant, however, at the beginning of the process, n_2' is small, also $|\Delta V| \gg V_{02}$, hence the second term is small. As the trapped ion density n_2' increases, $|\Delta V|$ also decreases, and the second term increases. At equilibrium the two terms must be equal, then $dN_2/dt = 0$, and Eq. (7) becomes,

$$\pi r^2 p P n_1 \left(\frac{2Ve}{m}\right)^{\frac{1}{2}} = 2\pi R n_2' \epsilon^{\Delta V/V_{02}} \left(\frac{V_{02}e}{2\pi M}\right)^{\frac{1}{2}}.$$

Cancelling terms, and solving for p, gives

$$p = \frac{R n_2'}{\pi^{\frac{1}{2}} r^2 P n_1} \left(\frac{m}{M}\frac{V_{02}}{V}\right)^{\frac{1}{2}} \epsilon^{\Delta V/V_{02}}. \tag{8}$$

This represents the equilibrium condition for a potential depression in the beam. It is evident that ΔV is a function of pressure, and that at a certain pressure, $\Delta V=0$. Under this condition also $n_2'=n_1$, and then

$$p_{(\Delta V=0)}=\frac{R}{\pi^{\frac{1}{2}}r^2P}\left(\frac{m}{M}\frac{V_{02}}{V}\right)^{\frac{1}{2}}. \tag{9}$$

Only at this pressure is there accurate space charge neutralization; at other pressures n_2' is not exactly equal to n_1. However, under practical conditions, such as are of interest here, ΔV is usually less than a volt, and then $n_2'\approx n_1$ for beam currents of several milliamperes or more. This arises from the fact that the potential depression is caused by the unneutralized charge n_1-n_2', and since these are both large numbers, n_1-n_2' may be large even though n_1/n_2' is nearly unity.

When the pressure exceeds the value given by Eq. (9), ΔV changes sign and the potential depression becomes a potential increase. This is due to the formation of ions at a rate greater than that at which they can escape without an assisting electric field.

In this case ions are no longer trapped, but electrons are, and thus it is now possible to set up equations for electrons, similar to Eqs. (5) and (6), representing the rates of electron formation and loss.

The rate of flow of electrons to the wall will be given by

$$\left(\frac{dN_1}{dt}\right)_1=2\pi Rn_1'\epsilon^{-\Delta V/V_{01}}\left(\frac{V_{01}e}{2\pi m}\right)^{\frac{1}{2}}. \tag{10}$$

The rate of formation of electrons in the beam will be

$$\left(\frac{dN_1}{dt}\right)_2=\pi r^2pPn_1\left(\frac{2Ve}{m}\right)^{\frac{1}{2}}. \tag{11}$$

Here n_1' is the density of the electron gas at the beam center, that is, the density of the trapped electrons. This does not include the beam electrons, which are not trapped, do not form a gas, and do not obey the Boltzmann law. Also V_{01} is the electron gas temperature in equivalent volts.

Equating Eqs. (10) and (11) and solving for p yields,

$$p=\frac{Rn_1'}{\pi^{\frac{1}{2}}r^2Pn_1}\left(\frac{V_{01}}{V}\right)^{\frac{1}{2}}\epsilon^{-\Delta V/V_{01}}. \tag{12}$$

This equation is valid only for potential rises in the beam, whereas Eq. (8) is valid only for potential depressions. For $\Delta V=0$ they should both be valid. Equating them, and putting $\Delta V=0$, yields,

$$\frac{n_1'}{n_2'}=\left(\frac{m}{M}\frac{V_{02}}{V_{01}}\right)^{\frac{1}{2}}. \tag{13}$$

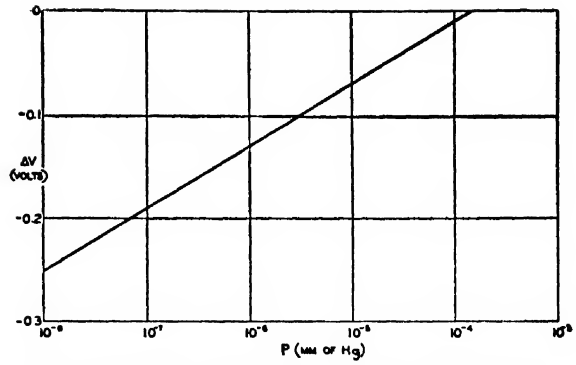


FIG. 5. Variation of potential depression with pressure.

If this is written

$$n_1'\left(\frac{V_{01}}{m}\right)^{\frac{1}{2}}=n_2'\left(\frac{V_{02}}{M}\right)^{\frac{1}{2}},$$

it is seen that this means that the electron and ion currents escaping from the beam, due solely to thermal motion, are equal, when $\Delta V=0$.

Equation (12) can not now be evaluated for other values of ΔV since n_1' is an unknown function of ΔV. Further work must be done to determine this function.

However, Eq. (8) for the potential depression case, can be evaluated, assuming $n_1=n_2'$. This is plotted in Fig. 5, for the following conditions: $R=0.5$ cm, $r=0.1$ cm, $P=7.7$, $M/m=52{,}000$, $V=300$ volts, and $V_{02}=0.026$ volt. These values correspond approximately to a set of experimental conditions to be discussed later. The values of R and r are averaged for a conical beam. P and M/m are taken for nitrogen. The ion gas temperature is assumed to be 300°K. These data give

$$p=1.5\times10^{-4}\epsilon^{\Delta V/0.026}.$$

It is seen from this that as the vacuum improves, the equilibrium potential depression increases in order to decrease the rate of ion escape from the trap. However for all practical vacuums the depression is very small compared to the beam voltage.

The build-up time of the ion density and the aperture current I_a may be computed from Eq. (7). However, this computation is complicated by the fact that the variation of ΔV during build-up is not known. Fortunately the term which contains ΔV does not become important until the final stages of build-up, when ions begin to escape from the trap. It is then responsible for the tail of the curves, as shown in Figs. 9 and 11. An approximation may be made by neglecting this term and using only the first term of Eq. (7). The values thus obtained will be too small, but will be sufficiently accurate for many purposes. If this is done, there results,

$$dn_2/dt=pPn_1v_1$$

since

$$N_2=\pi r^2n_2.$$

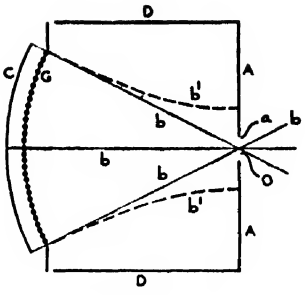


FIG. 6. Schematic diagram of ion-trapping gun.

The ion build-up time τ is the interval required for n_2 to equal n_1, i.e.,

$$\tau=\int_0^{\tau} dt=\int_0^{n_1}\frac{dn_2}{pP(V)n_1v_1}.$$

But $v_1=(2Ve/m)^{\frac{1}{2}}$, and also, if the number of ionized gas molecules is small compared to the total number of residual gas molecules in the beam-region, so that p can be considered constant during the ionization build-up period, it follows that

$$\tau=\frac{(m/2e)^{\frac{1}{2}}}{pP(V)n_1V^{\frac{1}{2}}}\int_0^{n_1}dn_2,$$

or

$$\tau=\frac{(m/2e)^{\frac{1}{2}}}{pP(V)V^{\frac{1}{2}}}.$$

If τ is measured in microseconds, p in mm of mercury, and $P(V)$ in number of ions formed per cm, per electron, at a pressure of 1 mm of mercury, this becomes

$$\tau=\frac{0.0169}{pP(V)V^{\frac{1}{2}}}. \qquad (14)$$

Thus τ varies inversely as the pressure, is independent of current, and is a function of voltage which can be specified only when $P(V)$ is known.

In the above treatment it is assumed that the number of residual molecules in the beam region is large compared to the number of ions required, that is, to the number of electrons in the beam. This condition is usually easily consistent with electron beam tube operating conditions.

IV. PULSE TECHNIQUE AND APPARATUS

The construction of the tubes used in the tests is shown schematically in Fig. 6. The cathode C comprises a segment of a spherical surface with its center of curvature at 0. Immediately in front of the cathode is the grid G, which also is a spherical segment having its center of curvature at 0. The aperture a, through which it is desired to pass current, is placed at 0. The disk A, cylinder D, and grid G form a shielded enclosure, which is field-free when G, A, and D are at the same potential. Under this condition, and in the absence of space charge, electrons will follow radial trajectories, as shown by lines b. But in the presence of space charge, divergence will occur, as illustrated by dashed lines b'.

A scale drawing of an actual tube is shown in Fig. 7. Parts are lettered to correspond to similar parts of Fig. 6. In addition there is shown here the focusing electrodes B, and the Faraday cage F to collect the aperture current I_a. The tests were made with a tube of demountable design in which the position of the electron gun and the Faraday cage could be adjusted by external knobs, working through metal bellows. This adjustment could be either lateral or longitudinal.

The greater part of the data obtained in this study was obtained by use of a pulse technique. It is believed that this is the first application of pulse methods to the investigation of electron beams. It has proved to be very superior to d.c. methods, eliminating several sources of error and yielding additional information. These advantages will be discussed below.

The circuit used is illustrated schematically in Fig. 8. The tube elements are cathode C, grid G, aperture disk A, and Faraday cage F. The pulser produced a square pulse of width T_2 variable from $\frac{1}{2}$ to 5 microseconds, and of spacing T_1 variable from 50 to 5000 microseconds. The pulse amplitude was adjustable up to 400 volts.

A typical oscillogram is sketched in Fig. 9, where grid voltage (above) and aperture current (below) are plotted against time. From a to b all electrodes have the same potential with respect to the cathode, and the current is constant. At b, a pulse is applied to the grid of sufficient amplitude to bring its potential down to zero. This stops the cathode emission current, and it also applies an ion-sweeping field between the grid and

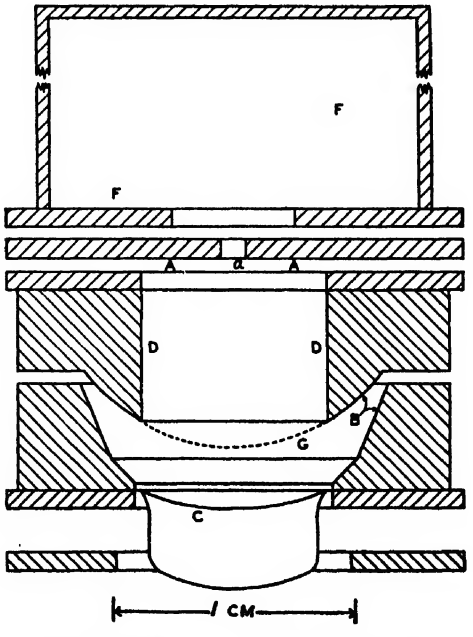


FIG. 7. Scale drawing of ion-trapping gun.

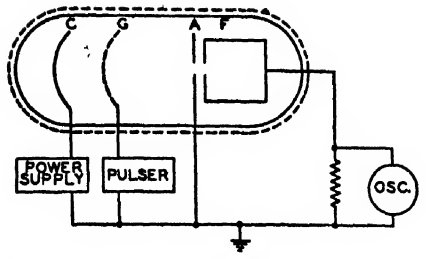


FIG. 8. Schematic diagram of tube and circuit.

the aperture for the duration of the pulse from c to d. At d the pulse is terminated and the grid-potential and the cathode emission current return to their initial values. The aperture current however rises only to e, which is less than the initial value. The reason for this is that the positive ions were swept out of the beam region during the application of the pulse, consequently at its termination, and the reformation of the beam, there is no space-charge neutralization. The current at e represents that passing through the aperture with the beam in a diverged state due to space charge. However, at point e ions again begin to be formed and trapped, hence there is an increase in current during the time interval from e to f. At the latter point the current has attained its original value, which represents that obtained with complete space-charge neutralization.

This method has several advantages over a dc method in which ions are swept out by applying a steady sweeping voltage on the grid. (1) Electrons maintain constant velocity between grid and aperture, as is required for comparison with existing theories. (2) There is no defocusing, or other distortion of the beam, by nonuniform fields in the beam region. (3) There are no complications caused by secondary electron currents. (4) An ion-sweeping field as large as desired may be used. (5) Data on ion build-up time are obtainable.

V. EXPERIMENTAL DATA

1. Space Charge Spread

The decrease in aperture current due to space charge spread may be obtained directly from oscillograms such as that of Fig. 9. The ratio of the current without SCN to that with, is the ratio of lengths ed/bc. The numerical value of the current at point b may be measured with a meter while the pulser is switched off. All other desired currents may likewise be metered except the current corresponding to point e. This is obtained from the ratio of ed/bc.

Measurements of this sort were made to determine the beam efficiency, $\delta = I_a/I$, as a function of beam current I, gas pressure p, and beam voltage V. Typical oscillograms are shown in Figs. 10–12.

The variation of I_a/I with beam current shows directly the effect of space charge in causing beam spread. Oscillograms for a series of beam currents are shown in Fig. 10. Here the beam voltage and pressure are substantially constant. As the current is varied, point e (see Fig. 9) changes, moving upward with decreasing current. This shows that the space charge spread is smaller for smaller currents. Data taken from such oscillograms are plotted in Fig. 13, where beam efficiency, I_a/I is plotted against beam current for three different beam voltages. The rapid drop in efficiency due to beam spread is apparent.

As pointed out above, these data may be presented in the form of a universal curve by plotting against $I/V^{\frac{1}{2}}$, instead of against I alone. This has been done in Fig. 3. The solid line has been computed from the Thompson-Headrick theory; the points have been obtained from ed/bc ratios on oscillograms such as those of Fig. 10. These data at different voltages show a systematic drop in efficiency as the voltage decreases. This will be shown later to be due to the effect of initial thermal velocities. Both voltage and pressure effects will be discussed later.

2. Space-Charge Neutralization

In Fig. 13 are shown also the values of I_a/I for space charge neutralized beams. These are substantially horizontal lines, except for a small upward rise for small currents. This indicates that there is little or no beam spread as the current increases in this case. Over the range of the measurements there appears to be no limitation of current due to space-charge spread. The tests were not usually extended beyond $I=55$ ma at $V=300$, because of excessive heating of the grid, and danger of destroying it. However, for a period of a few seconds, an aperture current $I_a=68$ ma was obtained at 400 volts with a beam efficiency $E=86.8$ percent. The lower voltage tests terminate at values of current which represent the space charge limited values of current between cathode and grid. This is not a fundamental or serious limitation since it can be varied by changing the cathode-grid spacing. (There is no SCN effect in the cathode-grid region since it is not field-free and ion trapping does not occur.) It is seen from these data that the maximum gain obtained in these tests from SCN corresponds to an increase in I_a of about 30 times at $I=55$.

It also is evident from Fig. 11 that the gain due to

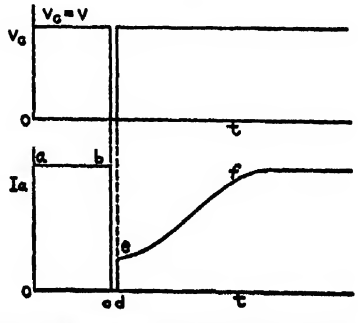


FIG. 9. Typical pulse shape showing applied grid voltage (above), and resulting aperture current (below).

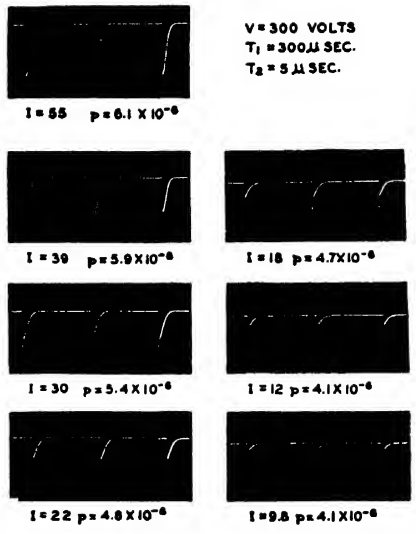


FIG. 10. Effect of current variation.

SCN is independent of pressure, since the ratio bc/ed does not vary with pressure. This is in agreement with the above theory.

3. Ion Build-Up Time and Pressure

The ion build-up time τ may be determined from oscillograms such as those of Fig. 9 by measuring the time difference between the points e and f. However, by using this method good agreement could not be expected when comparing the results with the simple theory used in deriving Eq. (14), because of some serious sources of error that cannot be overcome, namely: (1) The point f is very hard to localize on the scope picture because of the smooth transition to the equilibrium condition. (2) The tendency for the ions to escape from the beam region increases as the potential

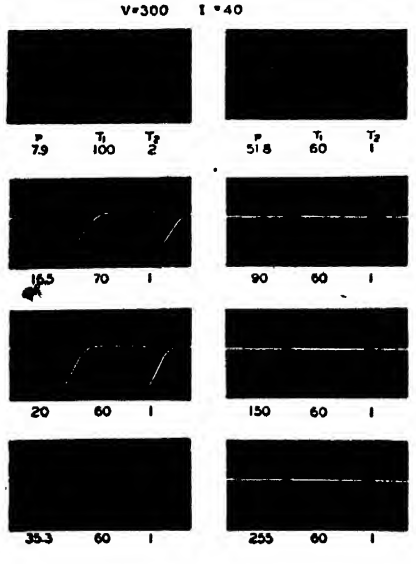


FIG. 11. Effect of pressure variation.

across the beam approaches the equilibrium state thus giving a longer ion build-up time than predicted by Eq. (14). These factors can be overcome by using another method, in which the build-up time is determined from the rate of formation of ions at the point e and from the knowledge of the number of ions required for full neutralization. The values thus obtained of course may differ somewhat from the practical ones, such as the time from e to f in Fig. 9, but they are consistent with the assumptions of Eq. (14) and are valid for comparison therewith. This method of obtaining τ has been used for all values given in this paper.

It is seen that, theoretically, according to Eq. (14), the build-up time varies inversely as pressure, inversely as $P(V)\ V$ and is independent of current. This latter independence can be seen to be substantially verified by the data of Fig. 10. No attempt at a more exact verification has been made.

The variation with pressure has been more completely checked. Pulse characteristics are shown in Fig. 11 for a range of pressures from 7.9×10^{-6} to 255×10^{-6} mm of mercury. These apply to a 40-milliampere, 300-volt beam. The decrease in build-up time is of an easily apparent magnitude in the photographs.

Values of τ obtained from such oscillograms are plotted against pressure in Fig. 14, on log-log paper. The solid lines are drawn with the theoretical slope of -1. The agreement lies within the experimental error. The 200-volt plot lies above the 300-volt one since the quantity $P(V)\ V$ of Eq. (14) is smaller in the former case. The values of $P(V)$ were taken for air.[14]

4. Variation of Beam Voltage

The build-up time is seen, from Eq. (14), to vary with beam voltage inversely as $P(V)\ V$. The function $P(V)$ in general has a value which rises to a maximum near a hundred volts and then drops slowly as V increases.[14] Multiplying by V shifts the maximum to a higher value. This function is inverted in the expression for τ, and results in a curve such as shown in Fig. 15, by the solid line. The experimental values, plotted as crosses, were obtained from oscillograms such as those of Fig. 12, by the method described in the preceding Section 3. Although the agreement is not quantitatively good, it is satisfactory in view of the assumptions involved. It is evident from known $P\ (V)$ functions that 300 volts represents an optimum voltage region, but that at higher voltages the build-up time would probably not be seriously larger.

At low beam voltages (below about 100 volts) it has been observed that the efficiency (I_a/I) drops very rapidly. This is illustrated in Fig. 16. It can be shown that this is caused by beam spread due to initial thermal velocities which become of importance when the beam velocity is small. To estimate this effect consider a

[14] See Engel and Steenbeck, *Elektrische Gasentladungen* (Berlin, 1932), p. 35.

small current, or an SCN case, so that space charge can be neglected. Assume a system such as shown in Fig. 17, where R_0 is the emitter radius, and r_2 the radius at the aperture, i.e., at the focal spot, due to initial velocities. Neglecting initial velocities, the focal spot would be a point, and the beam angle would be α, as shown.

For the system of Fig. 17, E. Ruska[15] has derived the expression

$$\frac{r_2}{R_0}=\frac{2\sin\beta}{\sin 2\alpha}\left(\frac{kT}{Ve}\right)^{\frac{1}{2}}.$$

To obtain the total spread put $\beta=\pi/2$. Also substitute the approximation,

$$\sin 2\alpha = 2R_0/z_0,$$

then

$$\frac{r_2}{R_0}=\frac{z_0}{R_0}\left(\frac{kT}{Ve}\right)^{\frac{1}{2}}, \quad \text{or} \quad \frac{r_2}{z_0}=\left(\frac{kT}{Ve}\right)^{\frac{1}{2}}. \tag{15}$$

Since the total current remains constant $\pi R_0^2 J_0 = \pi r_2^2 J$, or

$$J/J_0 = R_0^2/r_2^2, \tag{16}$$

where J_0 is the initial current density, and J the density at the focus.

From Eqs. (15) and (16),

$$J/J_0 = R_0^2/z_0^2(Ve/kT) = (Ve/kT)\tan^2\alpha.$$

For small α, $\tan\alpha \approx \sin\alpha$, so that this is in agreement with the expression of D. B. Langmuir.[16]

The fraction of current passing through an aperture of radius r_a is r_a^2/r_2^2, or from Eq. (15),

$$I_a/I = r_a^2/r_2^2 = r_a^2/z_0^2(Ve/kT). \tag{17}$$

This expression is valid, of course, only for $r_a \leq r_2$. Under this condition, and for a given geometry, the beam efficiency varies directly as V, the beam voltage.

This theoretical expression is plotted in Fig. 16 as a straight line through the origin. According to this simple theory the ratio I_a/I should rise linearly until it attains unity ($r_2=r_a$) after which it should remain constant. The departure of the experimental curve from the simple theory may be due in part to the neglect of the Maxwell-Boltzmann distribution in the theory. This distribution would cause a rounding off, instead of a sharp break, at $I_a/I=1$. Also there is spreading of the beam due to other factors, such as grid irregularities, which prevent the experimental value of I_a/I from reaching unity.

The slope of the line in the figure is drawn to equal that of the experimental curve at the origin and corresponds to a temperature of 1370°K. This is a little high, but the actual cathode temperature was not determined, and this value is close enough to indicate

[15] E. Ruska, Zeits. f. Phys. 83, 684 (1933).
[16] D. B. Langmuir, Proc. I.R.E. 25, 977 (1937).

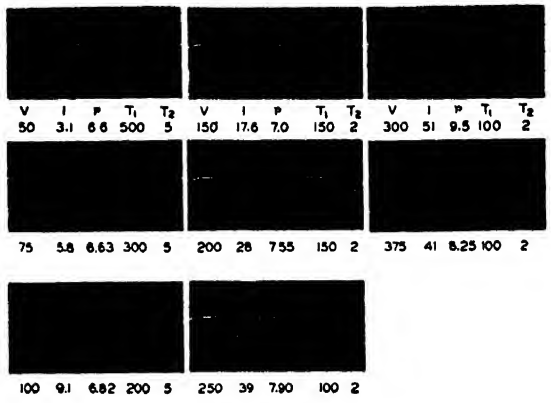


FIG. 12. Effect of voltage variation.

definitely that the drop in I_a/I is due to thermal velocities.

If the beam voltage is reduced below the ionization potential it is clear that the formation of ions will be affected and the neutralization of electron space charge may not be possible. Data taken in this region showed that as V increases from 9 to 10 volts there is an abrupt increase in I_a, corresponding, very likely, to the onset of ionization. However, as V is decreased from 10, the current drops smoothly until it reaches the original value at about 7 volts. Thus there is a hysteresis effect. This phenomenon is shown in the oscillogram of Fig. 18 wherein the horizontal scale was scanned by a 60 c.p.s. voltage, while the vertical represents I_a.

5. Comparative Results

Comparisons of SCN electron guns with several other types are given in Table I. The other types represent various previous attempts to obtain high current density, except, the 2K25 which is a widely-used commercial type, but not designed for particularly high density. The comparison is made on the basis of six quantities. As listed in order in the columns, they are: (1) total aperture current, (2) current density at the emitter, (3) current density at the aperture, (4) perve-

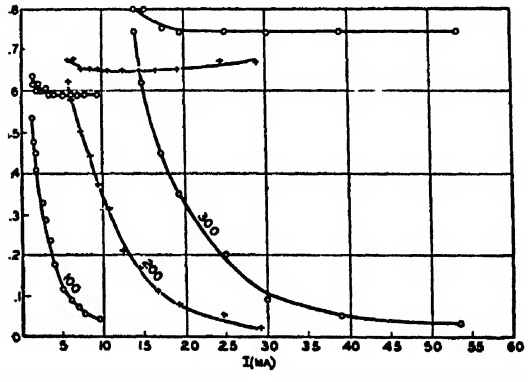


FIG. 13. Experimental beam efficiency curves with (above), and without ion trapping (below).

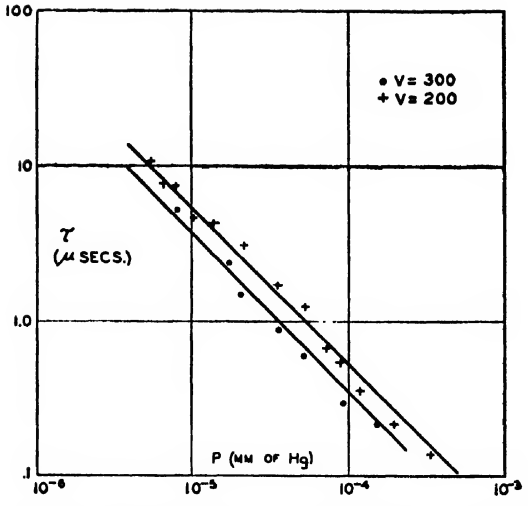


FIG. 14. Variation of ion build-up time with pressure.

ance$\times 10^6$, (5) concentration factor, or the ratio J/J_E, and (6) a figure of merit, F.

The figure of merit represents an attempt to indicate, in a manner independent of voltage and electrode size, the comparative ability of various guns to pass current through an aperture. To do this the Child-Langmuir law for parallel beams has been modified by introducing the factor F, i.e.,

$$I_a = 2.334\times 10^{-6}(V^{\frac{1}{2}}/x^2)a\cdot F,$$

where I_a is the current in amp./cm², V is the beam voltage, x is the distance from emitter to aperture in cm a is the aperture area in cm², and F is the figure of merit. For a parallel beam, $F=1$, and the greater the value of F the greater the aperture current compared to a parallel beam. Thus

$$F=\frac{I_a x^2 10^6}{2.334V^{\frac{1}{2}}a}. \tag{18}$$

For space charge limited beams this is independent of

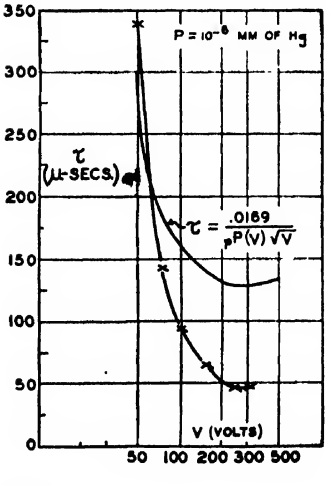


FIG. 15. Variation of ion build-up time with voltage, theoretical and experimental (X).

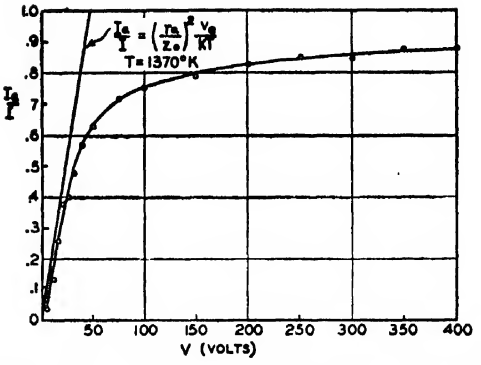


FIG. 16. Variation of beam efficiency with beam voltage. Current space charge limited between cathode and grid.

V, since I_a varies as $V^{\frac{1}{2}}$. Further, it is independent of gun size, since the ratio of x^2/a remains constant during scaling.

It is evident from the table that the SCN guns exceed all others in F value by a considerable factor. An important type of gun not listed in the table is one recently developed by O. Heil, for which very good results have been reported. This is not included, since published data are not yet available. It employs a very concave cathode, a wide-angle beam, and an optimized electrostatic focusing system.

VI. STABILITY

If reference is made to Fig. 11 it will be seen that at low pressures (10^{-5} mm of Hg) the oscilloscope trace is noise-free, but at higher pressures (90×10^{-6} and higher) there is a visible indication of noise. This is shown also in Fig. 19, which was taken with a beam voltage of 300, a current of 40 ma, and a pressure of 1.5×10^{-4} mm of Hg. It is noteworthy that the noise appears quite abruptly at a point where space charge neutralization apparently is completed. There is no evidence of it during the ion build-up period.

Reason to expect noise, and instability, under the observed conditions is found in the discussion of the equilibrium condition in Section III, and particularly in Eqs. (8) and (12). It was shown there that as the pressure increases, ΔV is at first negative, then passes through zero, and becomes positive. When ΔV is negative (potential depression) there is very nearly an exact equality of beam electron density and ion density. However, for ΔV positive, electron trapping starts,

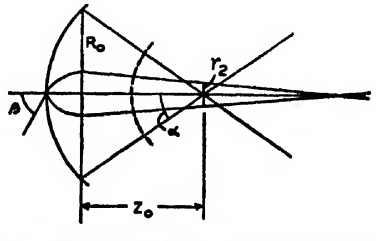


FIG. 17. Diagram showing beam spread due to thermal velocities in a structure with a focusing grid.

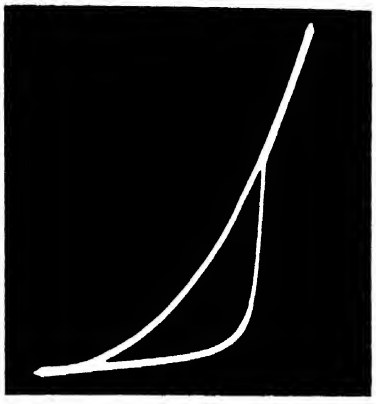


FIG. 18. Oscillogram showing variation of aperture current near ionization potential.

TABLE I. Comparative data on electron guns at 300 volts.

Type	I_a(ma)	$J_E\left[\frac{ma}{cm^2}\right]$	$J\left[\frac{ma}{cm^2}\right]$	Perveance $\times 10^6$	Conc. factor	F	Reference
Samuel	8.3	46.5	166	1.6	3.6	3.0	a
2K25	37.0	229.0	229	7.1	1.0	3.5	b
Lafferty	8.3	34.5	2200	1.7	64.0	38.0	c
SCN (1)	44.0	102.0	5700	8.5	56.0	470.0	
SCN (2)	49.0	113.0	6340	9.4	56.0	574.0	

[a] A. L. Samuel, Proc. I.R.E. 33, 233 (1945).
[b] J. R. Pierce and W. G. Shepherd, Bell Syst. Tech. J. 26, 460 (1947).
[c] J. N. Lafferty, thesis, University of Michigan, June, 1946.

and the electron and ion densities increase to some, at present, undetermined values. Furthermore, the temperature of the trapped electron gas in this case should be much higher than that of the trapped ion gas with ΔV negative. These conditions may well lead to noise production, and oscillations in the beam.

More specifically, noise has been found to be observable first at a pressure of approximately 10^{-4} mm of Hg. It rises to a maximum at 5×10^{-4}, and then remains constant to 10^{-3}, which is the upper limit of the experiments. It is seen that the threshold of 10^{-4} is in fair agreement with the theory, in that it is approximately at this value of p that ΔV changes from $-$ to $+$, as may be seen from Fig. 5.

FIG. 19. Oscillogram showing appearance of noise in SCN beam.

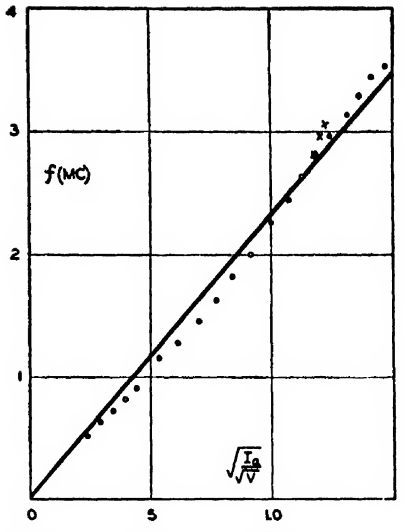


FIG. 20. Experimental beam oscillation frequency as a function of ion density.

Plasma ion oscillations also have been observed in the beam. These were found at the lowest pressures used, i.e., around 10^{-6}. They occurred up to $p=10^{-4}$ where they either ceased or were masked by the noise which begins at that pressure. It was not possible to determine which occurred.

For low pressures, the oscillation frequency was very distinct, and could be synchronized with an oscilloscope, or measured with a calibrated receiver. According to theory,[17] plasma ion oscillations have a frequency given by

$$f=\frac{1}{2\pi}\left(\frac{n_2e^2}{\epsilon_0M}\right)^{\frac{1}{2}}, \tag{19}$$

where M is the ion mass and n_2 the ion density. A measure of the ion density could be obtained from the current density at the aperture, and the beam voltage. Thus

$$f=K(I_a/V^{\frac{1}{2}})^{\frac{1}{2}}, \tag{20}$$

where K is a constant depending on ion mass. The agreement with this formula is shown in Fig. 20. The straight line represents Eq. (20) for a value of molecular weight of 41. This result seems high because the value for N_2 is 28, and O_2, 32, but since the identity of the trapped ions is unknown no definite conclusion can be drawn. However, the agreement is sufficiently good, so that there is little doubt that plasma ion oscillations are occurring.

No plasma electron oscillations were detected although they were sought over the range of from 80 to 1000 Mc, with the same tube as exhibited the ion oscillations.

[17] I. Langmuir and K. T. Compton, Rev. Mod. Phys. 2, 239 (1930).

Stress-Strain Properties of Natural Rubber under Biaxial Strain*

B. B. S. T. Boonstra
Delft, The Netherlands
(Received January 30, 1950)

Tube-shaped samples were stretched in a direction parallel to the tube axis and at the same time inflated to cause a tangential extension. In this way 600 percent axial and 300 percent tangential elongation could be attained simultaneously. The stresses were calculated from the pressure in the tube and the dynamometer readings.

The results are plotted as axial stress-strain curves at different tangential elongations ranging from 0–300 percent.

The curve for a higher tangential elongation lies at higher stresses than that for a lower tangential elongation; the differences are however larger than the theory predicts.

At 600 percent axial elongation stress relaxation was measured at different tangential elongations. It appears that the relaxation constant is independent of the tangential elongation.

INTRODUCTION

STRESS-STRAIN properties for rubber stretched in two directions have been reported by Ariano[1] and by Treloar.[2]

Both authors used rectangular test pieces and neither of them succeeded attaining high elongations because of the low tear resistance of rubber under biaxial stress.

High elongation was reached by the first author by inflating spherical samples, then however the strain in the majority of cases is about identical in the two directions. In this paper a system is described which allows stretching to high elongations in two directions independently of each other.

According to this method tube-shaped samples are stretched in the direction of the tube axis and at the same time inflated to cause a tangential elongation. The axial stress is partly measured on the dynamometer, the rest and the transverse stress is calculated from the air pressure in the tube. In this way stress-strain curves may be obtained for stresses parallel and at right angles to the tube axis as a function of the elongations in the two directions.

The theoretical equations extended to this case may then be checked. Besides, the decay of stress (axis-parallel) with time was measured and the influence of elongation perpendicular to this stress determined.

EXPERIMENTAL PROCEDURE AND RESULTS

Preparation of the Samples

Glass tubes of 16–25-mm diameter were repeatedly dipped in a latex dispersion of the following formula:

Rubber (in form of 40 percent latex)	100	
Colloidal sulfur	2	in form of 33 percent
Zinc oxide	2	dispersion with 0.2
Butyl zimate	1	percent Darvan
Latekoll (ammoniumpolyacrylate)	1	

After a layer of sufficient thickness was obtained (0.25–0.40 mm) the tubes were first dried at room temperature, then for 30 minutes at 80°C and finally vulcanized for 15–30 minutes at 110°C in live steam or hot air.

Dumb-bells punched out of these tubes showed the following properties:

Tensile strength	320–360 kg/cm²	
Elongation at break	750–800 percent	
Load at 300 percent elongation	17– 21 kg/cm²	
Hardness (shore *A*)	38	
Elasticity (shore *B*)	85	
Permanent set after 24 hours 200 percent elong.	measured after 1 hour	2.5 percent
	measured after 24 hour	0 percent

The rubber tubes are fastened on ribbed metal cylinders coated with a layer of vulcanized rubber (carbon black mix).

These cylinders can be attached to the dynamometer clamps; one of them has an air inlet connected to a manometer and a small hand operated air pump.

The tubes are elongated on a Schopper tensile tester (a) with air free to flow in and out, (b) inflated with air to the original diameter, (c) fastened on oversize clamp cylinders so that a constant transverse elongation can be maintained by pumping air into the tube.

The principle is illustrated in Fig. 1; some experimental values in Table I.

In calculating the stress in axis direction in case of definite transverse elongation, it has to be considered that the air pressure not only determines the tangential stress but also gives a contribution to the axial stress in form of pressure against the bottom and top plate of the cylinder.

Therefore the curves for inflated tubes do not begin at elongation zero but at a definite elongation determined by the air pressure necessary for the transverse elongation.

In Fig. 2 a number of curves, measured under different conditions is drawn, with axial stress, calculated on original cross section plotted against axial elongation.

The lowest curves are those measured without inflation, these practically coincide with the dumbbell curve;

* Communication No. 124. Rubber Foundation.

[1] R. Ariano, Rubber Chem. and Technol. 13, 92 (1940).

[2] R. L. G. Treloar, Trans. Faraday Soc. 40, 5 (1944); Proc. Phys. Soc. 60, 135 (1948).

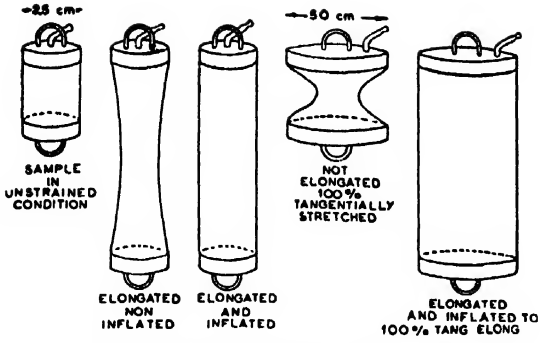


FIG. 1. Test samples for biaxial strain.

somewhat higher lies the curve for a tube which is not elongated sideways but inflated only to retain its original diameter. Considerably higher lie the curves for tubes with definite transverse elongations.

The pressure of the air, necessary to maintain a constant diameter of the tube at the different elongations is plotted as an average of several measurements against axial elongation in Fig. 3.

It will be observed in Fig. 3 that the pressure drops at the highest axial elongation for the 100 percent transverse elongation curve and also for the 300 percent elongation curve. Probably the rubber then has become so rigid that considerable changes in pressure are possible without much effect on the diameter of the tube.

The tangential stretching force on actual cross section is related to the pressure by the formula

$$\sigma = \frac{13.6 r \alpha_s}{1000(d/\alpha_s\alpha_z)} p \text{ kg/cm}^2 \text{ on actual cross section.} \quad (1)$$

Here r = original radius of the sample, mostly 12.5 mm, p = air pressure in cm Hg, α_z = axial extension ratio, α_s = transverse extension ratio, d = thickness of wall of the tube.

So at constant pressure and the same r, the actual stress in transverse direction is proportional to the

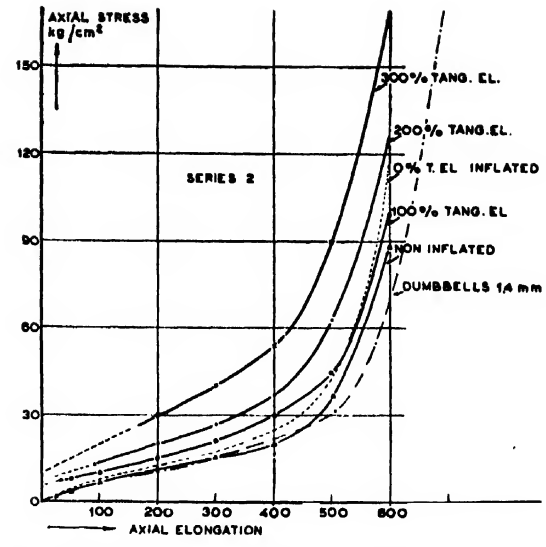


FIG. 2. Stress-strain curves at different tangential elongations.

square of the transverse elongation, (tr. el.)$^2 = \alpha_s^2$, this factor is four at 100 percent tr. el. and nine at 200 percent.

At the same pressures the actual stress at 200 percent tr. el. would be 9/4 of that at 100 percent tr. el. In reality however the stress ratio is lower, so it can happen that the pressure at 200 percent tr. el. is less than at 100 percent tr. el. (Fig. 3).

If the stress is plotted *vs.* the axial elongation, a set of curves is obtained as illustrated by Fig. 4.

In case the tangential tension is very small, for instance if the sample is not tangentially elongated, but only inflated to prevent contraction, the entity plotted in the figures was multiplied by ten; the stress in the direction of the axis is the actual value.

At 100 percent and higher transverse elongations the actual transverse stress has been plotted together with the stress in the direction of the axis, *vs.* axial elongation.

TABLE I. Some experimental values of tensile tests on inflated tube-samples (series 2) at 100 percent tangential elongation. Stresses are calculated in kg/cm² of the original cross section. r = original radius of the tube; α_z = Axial extension ratio; α_s = Tangential extension ratio; d = original thickness of the wall of the tube; p = air pressure in the tube in cm Hg.

Tangential elongation 100 percent		Axial elongation in percent 25	50	100	200	300	400	500	600
Average d = 0.033 cm	I Dynamometer axial stress kg/cm²	0.8	1.8	5.7	10.8	17.7	27.9	46.3	111.5
	II Pressure p in cm Hg	6.8	6.2	5.1	5.1	4.5	4.4	4.4	6.5
	III Axial stress calc. from $0.0272r/dp$ kg/cm² [a]	7.2	5.8	5.3	5.3	4.7	4.6	5.0	6.9
	I+III Total axial stress kg/cm²	8.0	8.6	11.0	16.1	22.5	32.5	51.3	118.4
	IV Tangential stress kg/cm² [b]	9.3	8.7	10.7	15.9	18.8	23.0	26.7	46.9

[a] General formula: $(13.6\pi r^2\alpha_s^2/2\pi rd1000)p$.
[b] General formula: $(13.6\alpha_s\alpha_s r/1000d)p$.

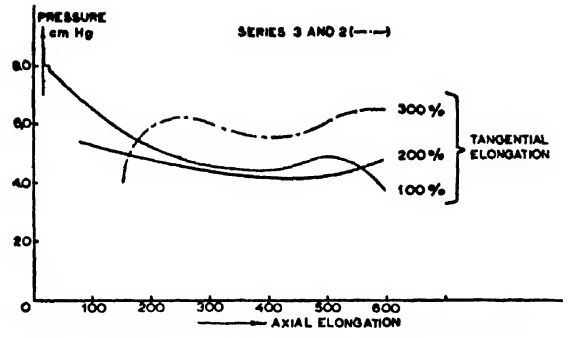


FIG. 3. Air pressure as a function of axial elongation.

It is clear from these figures that the tangential stress in all cases increases with axial elongation. The slope of this line increases with the transverse elongation.

Furthermore at a definite transverse elongation say 200 percent the tension must be equal to the stress parallel to the axis at an axial elongation of the same extent, in this case 200 percent. This is approximately true, however at 300 percent the referred stiffness of the tube caused somewhat higher values for the transverse stress.

RELAXATION

The decay with time of the stress parallel to the tube axis, was measured at the highest axial elongation i.e. 600 percent in the following way. The stress was measured at the moment of attaining the elongation of 600 percent and $\frac{1}{8}$—$\frac{1}{4}$—$\frac{1}{2}$—1—2—4—8—16 and 32 minutes after that.

A linear relationship is obtained if the stress is plotted *vs.* log time as is done for instance in Fig. 5.

The slope of this line $(F_{1/8}-F_{32}/-\log\frac{1}{8}+\log 32)$ is dependent on the original stress.

Instead of the original stress in our case the stress after one minute, F_1, is taken, as this value can be determined more accurately.

We calculated our relaxation constant from:

$$R_2=\frac{F_{1/4}-F_{16}}{6F_1}, \qquad (2)$$

taking the number 2 as a base of our logarithmic system.

The relaxation constant R_2 was calculated for measurements made with tubes under variable conditions. The results are put together in Table II.

It is apparent from these data that the tangential elongation does not exert much influence on the relaxation constant though the actual tension is much increased. Nor is there any influence of the nature of the mix in so far that soft sulfur-vulcanized rubber and thiuram vulcanized rubber show the same relaxation constant. However, if a mix with a greater amount of sulfur (e.g. five parts) is used the relaxation constant falls to considerably lower values. This holds almost independently of the time of cure of this latter compound.

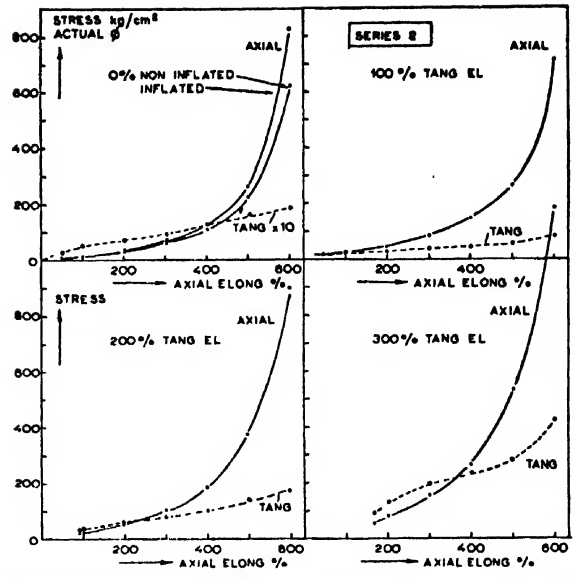


FIG. 4. Axial and tangential stress on actual cross section at biaxial elongation; rate of testing 30 cm per minute.

The relaxation constant is however sensitive to the axial elongation; at 500 percent the value is little more than half that at 600 percent elongation.

DISCUSSION

According to the now established statistical theory of elasticity[8] the elastic tension for unidirectional deformation is:

$$F=C[\alpha-(1/\alpha^2)]. \qquad (3)$$

If one of the cross dimensions (e.g. the circumference of the tube) is kept constant this formula changes into

$$F_c=C[\alpha-(1/\alpha^3)]. \qquad (4)$$

The following simple formulation of James and Guth's theory was kindly suggested by Ir. J. Bronkhorst of the Physical Department of the Rubber Foundation.

In Fig. 6, x, y, and z are dimensions of the rectangular sample. In

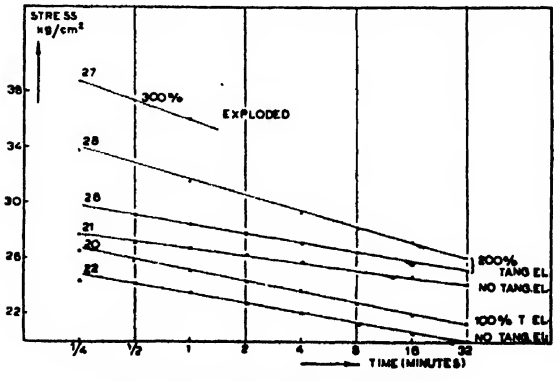


FIG. 5. Relaxation of stress with time.

[8] H. M. James and E. Guth, J. App. Phys. 15, 294 (1944); J. Chem. Phys. 11, 455 (1943).

unstrained condition $x=y=z=1$ cm. Then total molecular stress in x, y or z direction is $C=H$ (hydrostatic pressure in unstrained condition).

In a special strained condition x, y, z caused by forces F_x, F_y and F_z:

$F_x+Hyz=Cx=$ molecular tension
$F_y+Hxz=Cy$ $\quad x, y, z=1$ if the volume is considered constant.
$F_z+Hxy=Cz$

For unidirectional stress parallel to the x axis

$$F_y=F_z=0$$
$$Cy=Hxz$$
$$Cz=Hxy$$
$$F_x=Cx-Hyz=Cx-H/x$$
$$H=Cy/xz=C/x.$$
$$F_x=Cx-(C/x^2)$$

When z is kept constant by a force F_z, only $F_y=0$ (this is not exactly true as the inflation pressure represents a small negative value which is however neglected),

$$F_x+Hyz=Cx$$
$$Hxz=Cy$$
$$F_z+Hxy=Cz \quad H=Cy/xz=Cy^2=(C/x^2z^2).$$
$$F_x=Cx-Cy^2/x=C[x-(y^2/x)]=C[x-(1/z^2x^3)],$$
$$F_z=Cz-C(y^2/z)=C[z-(1/x^2z^3)].$$

Inflation to original diameter means $z=$ constant $=1$

$$F_x=C[x-(1/x^3)],$$
$$F_z=C[1-(1/x^2)].$$

Inflation to 100 percent transverse elongation ($z=2$):

$$F_x=C[x-(1/4x^3)],$$
$$F_z=C[2-(1/8x^3)].$$

Similarly at 200 percent transverse elongation:

$$\left.\begin{matrix}F_x=C[x-(1/9x^3)]\\F_z=C[3-(1/27x^3)]\end{matrix}\right\}.$$

At 300 percent transverse elongation:

$$\left.\begin{matrix}F_x=C[x-(1/16x^3)]\\F_z=C[4-(1/64x^3)]\end{matrix}\right\}.$$

For a definite elongation the second form is larger than the first one so it is to be expected that the stress-strain curve for a tube inflated to original cross section, will lie at higher stresses than the curve for the non-inflated tube. The difference should be relatively most pronounced at the lowest elongations.

The curves of Fig. 4 are not on a convenient scale to test these relations. In Table III several data are compared.

It appears from Table III that the axial stress of the inflated tube is relatively higher than would be expected from the theoretical formula. The stress of the non-inflated tubes at 100 percent and 200 percent corresponds with the less accurate values taken from dumbbells which amount to seven and ten kg/cm² respectively.

Similarly the ratio between axial stresses for non-inflated tubes and tubes inflated to 100 percent elongation is according to the theoretical derivation

$$\frac{\alpha-(1/\alpha^2)}{\alpha-(1/4\alpha^3)} \qquad (5)$$

In Table IV the calculated and experimental values are compared.

TABLE II. Relaxation constants at 600 percent elongation.

	Rate of stretching 15 cm/min.	30 cm/min.
Series II		
Non-inflated	...	$R_3=0.017$
Inflated to original diameter	...	$R_3=0.021$
Inflated to 100 percent tangential elongation	...	$R_3=0.019$
Inflated to 200 percent tangential elongation	...	$R_3=0.023$
Inflated to 300 percent tangential elongation	...	$R_3=0.021$
Series III		
Non-inflated	...	$R_3=0.017$
Inflated to 100 percent tangential elongation	$R_3=0.023$	$R_3=0.019$
Inflated to 200 percent tangential elongation	...	$R_3=0.018$
Series V		
Non-inflated	0.022	$R_3=0.018$
Inflated to original diameter	0.021	$R_3=0.019$
Inflated to 100 percent tangential elongation	0.019	$R_3=0.020$
Inflated to 200 percent tangential elongation	0.017	$R_3=0.018$
Inflated to 300 percent tangential elongation	...	$R_3=0.019$
Dumbbells cut out of latex tubes	...	$R_3=0.017$
Dumbbells cut out of sheets vulcanized with 1.75 parts of sulfur, 20′ at 142°C in a press	...	$R_3=0.020$[b]
As above but vulcanized with three parts of tuads[a] instead of sulfur +accelerator	...	$R_3=0.020$[b]
As above, vulcanized with five parts of sulfur+accelerator	...	$R_3=0.014$
Mixture with carbon black		$R_3=0.021$

[a] Tetramethyl thiuramdisulphide.
[b] These vulcanizates show a value of $R_3=$ about 0.013 at 500 percent axial elongation.

Though these figures are not very accurate they show a decided discrepancy with the calculated values; the experimental stresses being higher than those calculated.

This is even more obvious if we consider the transverse stress. Theory demands that the transverse stress, calculated on original cross section, shall remain practically constant at axial elongations above about 100 percent and under elongations where crystallinity disturbs the random configuration of the molecular network.

This is illustrated in Fig. 7 where the course of the

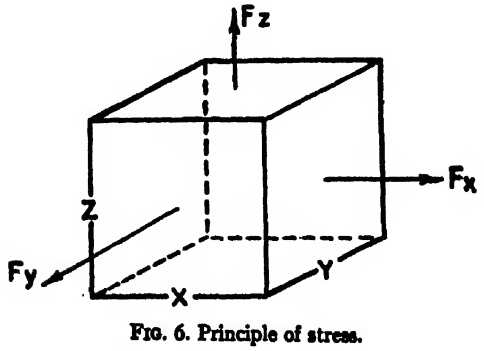


FIG. 6. Principle of stress.

TABLE III. Axial stresses on original cross section in kg/cm².

Series 2	Axial elongation 50 percent	100 percent	200 percent
Non-inflated	3.4	6.5	10.8
Inflated	4.4	7.9	12.3
Quotient	0.78	0.82	0.88
Quotient calculated from $\frac{[\alpha-(1/\alpha^2)]}{[\alpha-(1/\alpha^2)]}=$	0.87[5]	0.93[5]	0.97[5]

tangential stress is plotted schematically for a definite value of C (i.e., $C=3.3$). This value of C can be calculated from the stress-strain curve (non-inflated or dumbbell sample) as it is the axial stress where $\alpha-1/\alpha^2=1$ or $\alpha=$about 1.47 (47 percent elongation).

In our series 2 and 3 there is only approximate agreement at the very lowest elongations in axial or tangential direction. For instance at 0 percent tang. el., which means that the tube is inflated only to maintain its original diameter, the calculated curve only at the very beginning is approximately followed by the experiment. At 100 percent trans. el. however the first points up to about 100 percent axial elongation show some slight agreement, at higher elongations the difference increases strongly. It is to be expected that the constancy of volume assumed in theory will not hold exactly but it cannot explain the wide difference between experiment and theory.

Nor is it to be expected that neglecting of the small negative force in the third direction (the air pressure against the inner wall of the tube) would make much difference in the course of the theoretical curves.

So it seems that whereas there is reasonable agreement between experiment and theory for unidirectional stress this is no longer the case as soon as simultaneous elongation in a second direction is applied.

The high stresses above 300 percent axial elongation may point to the fact that no equilibrium is reached; if the tube decreases a little in diameter during axial stretching and afterwards is inflated to the correct diameter a higher pressure will be necessary than when both operations are accurately synchronized.

Treloar in his experiments about two-dimensional strain plotted the force per cm length of the extended sheet[4] (analogous to surface tension). As in his studies

TABLE IV. Calculated and experimental ratio of axial stress for non-inflated tubes and tubes inflated to 100 percent tangential elongation at various axial elongations.

Axial elongation	50 percent	100 percent	200 percent
Ratio calculated from.....(5)	0.74	0.89	0.97
Ratio measured in series 2	0.40	0.60	0.72
Ratio measured in series 3	0.52	0.71	0.80

[4] L. R. G. Treloar, Trans. Faraday Soc. 39, 241 (1943).

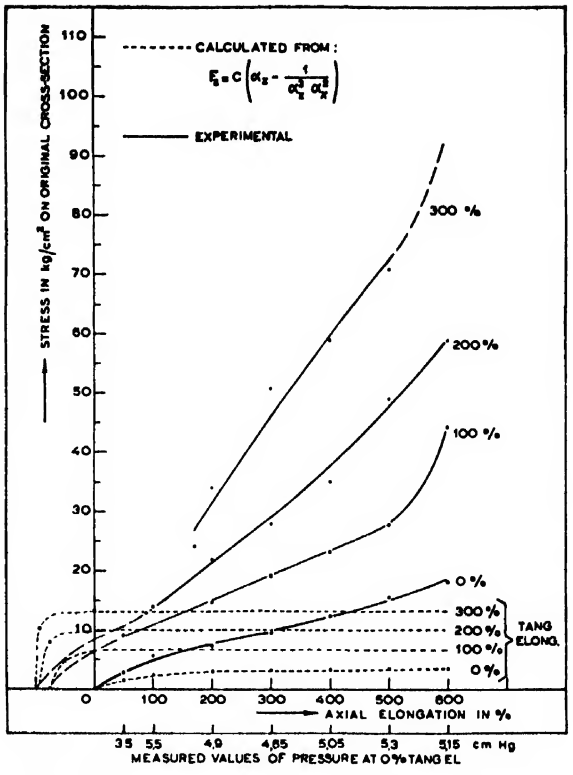


FIG. 7. Tangential stress as a function of axial elongation.

the strain in two dimensions was identical, the force per cm length of the sheet is

$$t=S_z/X^2=F_z/X=C\left(1-\frac{1}{X^6}\right)=C(1-\gamma^3), \quad (6)$$

when γ is the ratio of the final to the original thickness of the sheet and $X=Z$ the extension ratio in a direction parallel to the surface.

Treloar could show that the experimental tension force t at different extensions coincided with the theoretically derived equation

$$t=C(1-\gamma^3) \quad \text{for} \quad C=\text{about 4 kg/cm}^2$$

up to about 200 percent elongation ($\gamma=\frac{1}{3}$).

In our case the strain in the two principal directions is generally not the same but at every tangential elongation we have one moment where axial and tangential elongation are identical. In series V we have these cases at 56 percent, 100 percent, 200 percent, 275 percent and 300 percent elongation. At these points of identity of the two principal strains the stresses in the two perpendicular directions must be identical too; this appears to be the case with a deviation of about the experimental error of the stress measurements.

To compare our results with those of Treloar we have plotted the force t for these cases in Fig. 8 and it is evident from this figure that they agree fairly well with

theoretically calculated values for $C=5.7$ up to 200 percent elongation. This is shown in Table V.

So far there is agreement with theory as had been found by Treloar except that in our case the value of C is 5.7 against 4.0 in Treloar's experiments.

This however may be due to the different types of mixture used.**

In case the strains in the two perpendicular directions across the sheet are not identical we have two principal stresses and following again the method of plotting the entity t (stress per cm length of the sheet) we easily derive:

$$t_x=\frac{S_x}{XZ}=\frac{F_x}{Z}=C\left(\frac{X}{Z}-\frac{1}{X^3Z^3}\right)$$

and

$$t_z=\frac{S_z}{XZ}=\frac{F_z}{X}=C\left(\frac{Z}{X}-\frac{1}{X^3Z^3}\right)$$

for $X/Z=1$ this case simplifies again to the first problem. If we take however the strain situations, where X/Z is for instance $=2$ we would expect to find a similar curve as in Fig. 8 with the same value of C.

What actually is found Table VI illustrates. From Table VI it is evident that the C values do not agree with the C found for homogeneous two-dimensional strain and differ even when calculated from the axial and tangential stress respectively. So here also the deviations from theory are very pronounced. Mathematically these deviations may be accounted for partly by the treatment of Mooney[5] and more completely by that of Rivlin.[6]

The formula governing the relaxation which has been used here is

$$R=\frac{Ft_1-Ft_2}{Ft_1/(\log t_2-\log t_1)}; \qquad (7)$$

this expresses a constant decrease of stress in all time

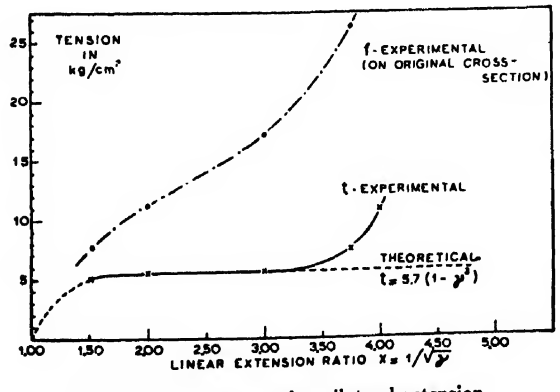


FIG. 8. Two-dimensional equilateral extension.

** For latex-rubber T. mentions a value of $C=6.3$.

[5] M. Mooney, J. App. Phys. 11, 582 (1940).

[6] R. S. Rivlin, Phil. Trans. A, 240, 459, 509 (1948); 241, 379 (1948); 242, 173 (1949); Proc. Roy. Soc. A, 195, 463 (1949).

TABLE V. Forces at various equilateral elongations and the values of "C" calculated from these forces.

Elongation	F_x/X	"C"
56 percent	7.9/1.56=5.1	5.7
100 percent	11.2/2=5.6	5.6
200 percent	17.2/3=5.7	5.7
275 percent	26.5/3.75=7.1	7.1
300 percent	46/4=11.5	11.5

periods t_2-t_1, where $t_2=2t_1$ and t is counted from the moment the elongation was stopped. It is an approximation of the general formula of Tobolsky and Eyring.[7] It indicates that the decay of stress is the result of slippage of chain molecules along each other.

According to Flory[8] the relaxation results for the greater part from terminal chain slippage. This means that a rubber with longer primary molecules (before vulcanization) will have a lower relaxation constant than a rubber which has been thoroughly broken down to small molecular chains on the mixing mill.

If we compare two rubbers that have undergone the same milling treatment, so that we may assume about equal average molecular weight, but that have different degree of vulcanization, then the one with the highest cure will show the lower relaxation constant as the latter has the highest percentage of cross links and relatively the lowest number of terminal chains.

This is in agreement with the observation that the vulcanizate with five parts of sulfur shows a constant $R=0.014$ whereas the mixes vulcanized with 1.75 parts of sulfur and with thiuram only, show a value of $R=0.020$.

At these high elongations crystallization plays a part in stress-relaxation. The crystallites begin to form at about 250–300 percent elongation; this means that these first crystallites have a comparably low degree of

TABLE VI. Values of "C" calculated from axial and tangential stresses with $X/Z=2$.

Series V	Fx/Z	C calculated from Fx/Z	Fs/X	C calculated from Fs/X
Ax. el. 100 percent / Tang. el. 0 percent	7.0– 7.5	3.7–4	2.6–3.2	6.9–8.6
Ax. el. 200 percent / Tang. el. 50 percent	9.1	4.6	3.6	7.2
Ax. el. 300 percent / Tang. el. 100 percent	10.1–11.8	5.1–5.9	3.9–4.0	8.0
Ax. el. 400 percent / Tang. el. 150 percent	14.2–15.8	7.2–7.9		
Ax. el. 500 percent / Tang. el. 200 percent	17.2–18.5	8.6–9.3	4.0–5.0	8–10

[7] A. V. Tobolsky and H. Eyring, J. Chem. Phys. 11, 125 (1943). The simplified formula was experimentally determined by P. Phillips, Proc. Phys. Soc. (London) 19, 491 (1905).

[8] P. J. Flory, Ind. Eng. Chem. 38, 417 (1946).

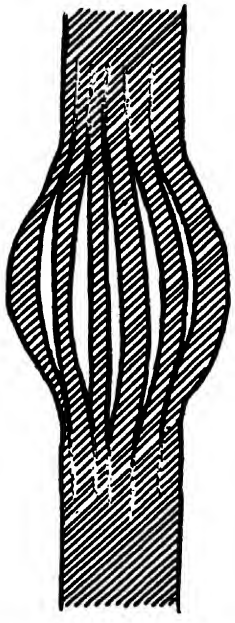


FIG. 9. Effect of inflation on a stretched sample.

orientation, lower than those formed at 400-500 percent elongation and these again are less orientated than crystals formed at 500–600 percent elongation.

In the latter state there are many crystallites that were in equilibrium with the surrounding amorphous mass at elongations much lower than 600 percent.

These strive to reach an equilibrium at the now (600 percent el.) prevailing conditions by re-crystallization and detaching of terminal chains from crystallites. It is probable that such a re-crystallization is a rather slow process as compared with the slippage of terminal chains but it is difficult to distinguish between them..

Mooney *et al.*[9] besides, report indications of a relaxation mechanism of a higher rate than that found in normal analysis; its existence was established on Shore Durometer readings $\frac{1}{10}$ sec. after stretching.

In our case however we would expect that transverse strain would severely hamper orientation and crystallization in axis parallel direction and diminish its contribution to stress-relaxation. The effect on crystallization can be experimentally shown by first strongly elongating one of our tube-samples and then inflating it. Very high air pressure is necessary to cause the first transverse elongation which then does not turn out homogeneously but occurs by a definite tearing of the opaque crystallized material into streaks of thinner and clear amorphous areas and remaining crystallized fiber-like bundles. This situation is schematically drawn in Fig. 9.

There has not been found definite influence of the transverse strain on the axial relaxation. It may be that the measurements, which are technically difficult to perform and absorb much time, have not been sufficiently accurate to distinguish the differences but then the effects could not be very pronounced.

CONCLUSIONS

(1) The stress-strain curves of vulcanized rubber held at definite tangential elongations during stretching differ more from the simple unidirectional-extension-stress curve than the theory predicts. The differences are larger at higher tangential elongations.

(2) The relaxation constant determined at 600 percent elongation is practically unaffected by tangential elongations up to 300 percent. The investigations are part of the work on fundamental properties of rubber of the Research Department of the Rubber-Stichting at Delft.

[9] Mooney, Wolstenhome, and Villars, J. App. Phys. 15, 324 (1944).

On the Endurance Limit of a Round Bar with Longitudinal Grooves

H. Ōkubo
Institute of High Speed Mechanics, Tōhoku University, Sendai, Japan
(Received June 2, 1950)

In a previous paper, we reached a consideration that the endurance limit of a material should be determined by the amount of the mean stress in a certain area around a point, instead of the stress at the point, and the extent of the area is determined for several kind of metals, by using the results of tension-compression fatigue tests.

In this paper, the extent of area where the mean value of stress to be taken, is determined for a mild steel in another way, viz., by a torsion fatigue test, and we reached the conclusion that the extent of area chiefly depends on the kind of metal and independent of the course of the experiment.

IN our previous study, the influence of the diameterical hole on the endurance limit of round bars has been investigated.[1] The theory of elasticity shows that the stress concentration around a diameterical hole of a round bar under tension, increases as the diameter of hole decreases and when the diameter of hole decreases infinitely, the stress concentration factor approaches 3.[2] While, according to our experimental results for several kinds of metals, with decrease in the diameter of the hole, the endurance limit increases, when the diameter of the hole is very small. The theory fails to explain the experimental results. The discrepancy between the theoretical and the experimental results arises from a fact that the theory assumes the material to be perfectly homogeneous and isotropic, while the metals in practical use are neither homogeneous nor isotropic. To correct the theory so as to meet the experimental results, we reached the consideration that when we determine the endurance limit, we must consider the mean stress in a certain area around the point, instead of the stress at a point. To simplify the calculation, we took up the mean value of the stress in a certain line, in place of the mean value in the area. The magnitude of the distance h where the mean value is taken, was considered to depend chiefly on the size of crystals of the metal and so it is a constant, determination of which is pending on the kind of metals. Under the above considerations, the values of h for some metals, determined by using our previous fatigue data, were

Mild steel	0.23 mm
Special steel	0.15 mm
Super-duralumin	0.13 mm

Our experiments for these metals were done by means of tension-compression tests using Haigh's testing machine.

In this paper, we shall determine the constant h for a mild steel in another way, *viz.*, by means of a torsion fatigue test, using specimens of round bars with longitudinal grooves as shown in Fig. 1. The specimens are cut out very carefully and the dimensions of grooves checked by a plate gauge which is measured accurately by a Zeiss comparator.

The mechanical properties of the steel are given in Table I.

The testing machine, used in our present experiment is of mechanical type. The specimen is placed in series with a weigh-bar, whose stress-strain relation has been previously obtained by statical loading experiments. The torque acting on the specimen is estimated by observing the twist of the weigh-bar and its twist is measured optically by means of a mirror placed at its moving end.

The scale is so graduated as to allow the direct reading of the maximum shearing stress for specimens of the diameter 10 mm. The fatigue data obtained in connection with the above mentioned program are given in Tables II and III. Figure 2 shows a fractured specimen with longitudinal grooves. The endurance limit of the steel is 11.5 Kg/mm².[3]

The fatigue cracks, appeared on the bottoms of grooves of specimens for different loadings, are shown in Fig. 3. When the magnitude of the reversed stress slightly passes over the endurance limit, very small cracks, as shown in Fig. 3(a), begin to appear locally on the bottom of the groove. According as the stress increases, the small cracks appear at every part of the bottom and join together, and then some of the enlarged cracks spread along the bottom of the groove (Fig. 3(b)), and finally the specimen is broken. We shall define the endurance limit for the specimens with grooves, by the

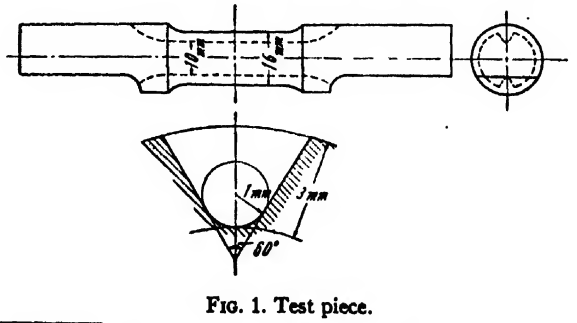


FIG. 1. Test piece.

[1] K. Minamiōzi and H. Ōkubo, J. Franklin Inst. 249, 49 (1950).
[2] R. E. Peterson, J. App. Mech. (ASME), p. A-15 (1936).
[3] For the endurance limit of specimens without grooves, the maximum stress is taken, since the variation in the interval h is very small.

FIG. 2.

limiting value of the stress, below which cracks will not occur. Then, the reading corresponding to the endurance limit for the specimens with grooves will be determined from Table III, as 17.3 Kg/mm².

Next, we shall make some calculations to find the stress distribution in the specimen corresponding to the reading.

Let us consider a circular cylinder of the cross section as shown in Fig. 4, and take the axis of z to be parallel to the generators of the cylinder. Then the stresses which do not vanish are

$$X_z=\mu\tau\left(\frac{\partial\phi}{\partial y}-y\right),\quad Y_z=-\mu\tau\left(\frac{\partial\phi}{\partial x}-x\right),\qquad(1)$$

where μ and τ represent the modulus of rigidity and the twist respectively. The function ϕ satisfies the equation

$$\partial^2\phi/\partial x^2+\partial^2\phi/\partial y^2=0,\qquad(2)$$

at all points within the bounding curve of the cross section, and at the boundary the condition

$$\phi-\tfrac{1}{2}(x^2+y^2)=\text{const.}\qquad(3)$$

We shall consider a function ϕ in the form[4]

$$\phi=\frac{(1-r^4)A_0}{1+r^4-2r^2\cos2\theta}+A_1\left\{\frac{1-a_1^2r^2\cos2\theta}{1+a_1^4r^4-2a_1^2r^2\cos2\theta}-\frac{r^4-a_1^2r^2\cos2\theta}{a_1^4+r^4-2a_1^2r^2\cos2\theta}\right\}.\qquad(4)$$

TABLE I.

Yield strength Kg/mm²	Ultimate strength Kg/mm²	True breaking strength Kg/mm²	Elongation percent	Reduction of area percent
25.5 (upper) 22.3 (lower)	38.2	83.0	36.6	67.9

TABLE II. Specimens of 10-mm diameter without grooves.

Specimen No.	Range of stress Kg/mm²	No. of repetitions	Remarks
1	±13.0	0.106×10^7	broken
2	±12.6	0.392×10^7	broken
3	±12.0	0.596×10^7	broken
4	*±11.5*	1.003×10^7	unbroken

[4] H. Ōkubo, "On the torsion of a shaft with keyways," Quart. J. Mech. App. Math. (to be published).

TABLE III. Specimens of 16-mm diameter with longitudinal grooves, as shown in Fig. 1.

Specimen No.	Reading of the nominal range of stress Kg/mm²	No. of repetitions	Remarks
1	±43.2	0.037×10^7	broken
2	±36.3	0.292×10^7	broken
3	±33.1	0.724×10^7	broken
4	*±32.7*	1.097×10^7	unbroken, cracked
5	±25.5	1.043×10^7	unbroken, cracked
6	±22.8	1.000×10^7	unbroken, cracked
7	±22.1	1.320×10^7	unbroken, cracked
8	±19.2	1.002×10^7	unbroken, cracked
9	±18.3	1.100×10^7	unbroken, cracked
10	±17.8	1.019×10^7	unbroken, cracked
11	*±17.3*	1.000×10^7	uncracked

The above function satisfies the Eq. (2), and is singular at the two points ($\theta=0,\ \pi$) on the bounding circle and at the two points ($r=a_1,\ \theta=0,\ \pi$) within the circle (a_1 being assumed to be smaller than unity).

From Eqs. (3) and (4), the equation which represents the boundary of the cross section is

$$\frac{(1-r^4)A_0}{1+r^4-2r^2\cos2\theta}+A_1\left\{\frac{1-a_1^2r^2\cos2\theta}{1+a_1^4r^4-2a_1^2r^2\cos2\theta}-\frac{r^4-a_1^2r^2\cos2\theta}{a_1^4+r^4-2a_1^2r^2\cos2\theta}\right\}-\tfrac{1}{2}r^2=-\tfrac{1}{2}.\qquad(5)$$

Equation (5) represents a circle of radius unity and the two curves represented by the equation:

$$\frac{A_0}{1+r^4-2r^2\cos2\theta}+\frac{A_1\{a_1^4(1+r^4)-a_1^2(1+a_1^4)r^2\cos2\theta\}}{(1+a_1^4r^4-2a_1^2r^2\cos2\theta)(a_1^4+r^4-2a_1^2r^2\cos2\theta)}+\frac{1}{2(1+r^2)}=0.\qquad(6)$$

When the curve represented by Eq. (6), agrees almost with the shape of the given notch, then we can use the Eq. (4), as an approximate solution for the problem.

To determine the unknown constants A_0 and A_1 in Eq. (4), we shall assume that the curve represented by Eq. (6) passes through the two points ($r=0.625$, $\theta=0°$ and $r=0.81623$, $\theta=12°10'$) on the periphery of the notch, then assuming $a_1=0.7562$, we have

$$A_0=-0.065507,\quad A_1=-0.045069.$$

If we denote the radius of curvature at the bottom of the notch by ρ, then

$$\rho=\frac{b^2}{b-(d^2r/d\theta^2)_0},\qquad(7)$$

where

$$\left(\frac{d^2\chi}{d\theta^2}\right)_0=-\frac{[8A_0b^2(1+b^2)/(1-b^2)^3]+4A_1a_1^2b^2\{[1+a_1^2b^2/(1-a_1^2b^2)^3]+[a_1^2+b^2/(a_1^2-b^2)^3]\}}{b-[4A_0b/(1-b^2)^2]-2A_1a_1^2b\{[1/(1-a_1^2b^2)^2]+[1/(a_1^2-b^2)^2]\}},$$

and b is the distance from the center of the cross section to the bottom of a notch.

Putting $b=0.625$, it becomes $\rho=0.1250$. The curve represented by Eq. (6), is shown by a dotted line in Fig. 4, which agrees practically with the given shape of the notch as is shown in the figure. Accordingly, using the solution given by Eq. (4), we shall proceed with our further calculations.

The stress on the x axis is calculated from Eq. (1), as

$$Y_x=\mu\tau\left\{r-\frac{4A_0r}{(1-r^2)^2}-2A_1a_1^2r\left(\frac{1}{(1-a_1^2r^2)^2}+\frac{1}{(a_1^2-r^2)^2}\right)\right\}, \quad (8)$$

and so the mean value of the stress from the bottom of the notch to a distance h along the x axis is

$$\bar{Y}_x=\frac{1}{h}\int_{b-h}^{b}Y_x dr=\frac{\mu\tau}{h}\left|\frac{r^2}{2}-\frac{2A_0}{1-r^2}-A_1\left(\frac{1}{1-a_1^2r^2}+\frac{a_1^2}{a_1^2-r^2}\right)\right|_{b-h}^{b}. \quad (9)$$

The value of $\bar{Y}_x$ varies with the magnitude of the distance h, and the relation between $\bar{Y}_x$ and h, calculated by Eq. (9), is shown in Fig. 5. The torque acting on the cylinder is[5]

$$T=\mu\tau\left(I-\oint\psi\frac{\partial\phi}{\partial s}ds\right), \quad (10)$$

where I is the polar moment of inertia of the cross section about z axis, ψ is a function conjugate with ϕ and

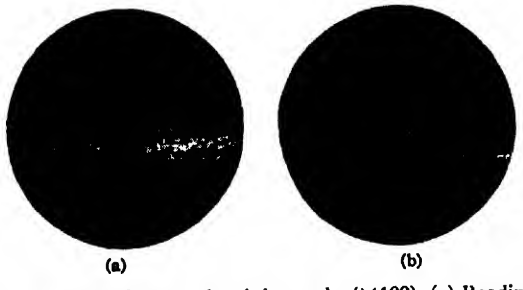


FIG. 3. Micro-photographs of the cracks (×100). (a) Reading of the nominal range of stress, ±17.8 Kg/mm². No. of repetitions $=1.019\times10^7$. (b) Reading of the nominal range of stress, 25.5 Kg/mm². No. of repetitions $=1.043\times10^7$.

[5] H. Love, *The Mathematical Theory of Elasticity* (1927) 4th Ed., p. 312.

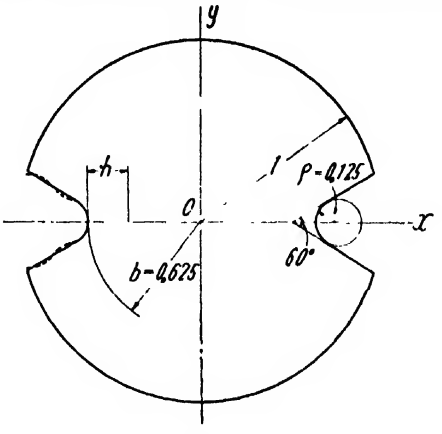


FIG. 4.

given by

$$\psi=-\frac{2A_0r^2\sin2\theta}{1+r^4-2r^2\cos2\theta}-A_1a_1^2r^2\sin2\theta\left\{\frac{1}{1+a_1^4r^4-2a_1^2r^2\cos2\theta}+\frac{1}{a_1^4+r^4-2a_1^2r^2\cos2\theta}\right\},$$

and the integral is taken around the bounding curve of the cross section.

For the section, shown in Fig. 4,

$$I=1.398 \quad \text{and} \quad \oint\psi\frac{\partial\phi}{\partial s}ds=0.311,$$

and hence

$$T=1.087\mu\tau.$$

The torque T_0 acting on a cylinder of radius 0.625 without grooves, when twisted by the amount τ_0 is

$$T_0=0.2397\mu\tau_0.$$

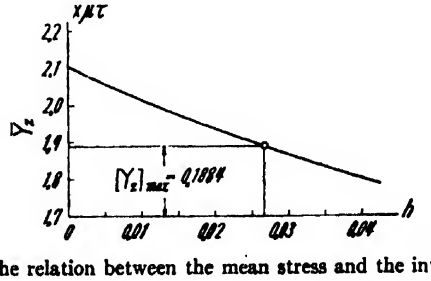


FIG. 5. The relation between the mean stress and the interval h.

Hence, we have

$$T_0/T=0.2205(\tau_0/\tau).$$

According to our fatigue data for a cylinder of radius 5 mm, without grooves, the endurance limit of the metal is 11.5 Kg/mm². While the reading corresponding to the endurance limit for a cylinder of radius 8 mm with grooves is 17.3 Kg/mm². The reading does not denote directly the maximum stress but represents the value proportional to the torque acting on the specimen, when the dimensions of the specimen differ from the standard type (a cylinder of radius 5 mm without grooves), subsequently

$$T_0/T=11.5/17.3,$$

and hence

$$0.2205(\tau_0/\tau)=11.5/17.3,$$

or

$$\tau_0=3.015\tau.$$

The maximum stress arises in a cylinder of radius 0.625, when twisted by the amount τ_0 is

$$[Y_z]_{max}=0.625\mu\tau_0=1.884\mu\tau.$$

The endurance limit of a metal, however, is a constant, which is independent of the form or the size of the specimen. Subsequently, $[Y_z]_{max}$ must be equal to Y_z, by virtue of our previous assumption that for the determination of the endurance limit of specimens, the mean value of stress in the interval h must be taken up instead of the maximum stress.

If we determine h from our fatigue data, so as to satisfy the above mentioned condition (see Fig. 5), it becomes

$$h=0.0265,$$

or

$$h=0.21 \text{ mm}.$$

The value of h for a mild steel, previously obtained by using our tension-compression fatigue data is

$$h=0.23 \text{ mm},$$

and the results almost agree in two cases.

From this result, we can conclude that the value of h depends only on the kind of metal, and independent of the course of the experiment.

ACKNOWLEDGMENT

A grant in aid for scientific research of the Education Ministry of Japan has been given to the present study.

The author also takes this opportunity to thank his assistants, T. Uematsu and E. Itagaki, for their help in this study.

The Magnetic Field of a Plane Circular Loop

C. L. BARTBERGER

Naval Air Experimental Station, Philadelphia, Pennsylvania

(Received March 9, 1950)

The axial and radial components of the magnetic field of a plane circular loop are expressed in terms of cylindrical coordinates. The expressions involve two integrals which are related to certain of the complete elliptic integrals. Tables of values of these integrals are presented. Interpolation in these tables facilitates rapid calculation of the field components.

INTRODUCTION

THE problem of calculating the radial and axial components of the magnetic field produced by the current in a plane circular loop of wire is well known and has been treated in numerous textbooks.[1,2] Recently Blewett,[3] making use of the solution in terms of cylindrical coordinates and complete elliptic integrals, has computed and tabulated the values of the field components of numerous points in the vicinity of the loop. Such a tabulation, however, has the difficulty of being two-dimensional, since the field is a function of both the axial and radial coordinates. It is therefore not practical to tabulate the field at a sufficient number of points to permit easy interpolation.

A few years ago the need arose to determine the magnetic field produced by a pair of Helmholtz coils which were used for testing aircraft compasses. The solution was carried out in cylindrical coordinates by a slightly different method from that used by Blewett. The resulting expressions for the field components are relatively simple and involve two integrals which are related to certain of the complete elliptic integrals. Once these integrals have been evaluated, the computation of the field components to high accuracy becomes a simple matter. It is the purpose of this paper

[1] J. H. Jeans, *The Mathematical Theory of Electricity and Magnetism*, 5th Edition (Cambridge University Press, London, 1925), pp. 431, 432.

[2] W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939), pp. 267 and 270.

[3] J. P. Blewett, "Magnetic field configurations due to air core coils," J. App. Phys. 18, 968 (1947).

to publish tables of values of the two integrals together with formulas for their use. Values are given for sufficiently small increments of the argument to permit linear interpolation throughout the greater part of the tables.

MAGNETIC FIELD COMPONENTS DUE TO CURRENT IN A PLANE CIRCULAR LOOP

If z and ρ represent axial and radial coordinates measured from the center of the loop, the axial and radial components of the magnetic field can be expressed as follows:

$$H_z=(H_0/s^3)[I_1-(\rho/a)I_2] \tag{1}$$

$$H_\rho=(H_0z/s^3a)I_2, \tag{2}$$

where a = radius of the loop, $s^2=(a^2+z^2+\rho^2)/a^2$, and H_0 = field at center of loop. The integrals I_1 and I_2 are

$$I_1=\frac{1}{\pi}\int_0^\pi \frac{d\theta}{(1-b\cos\theta)^{\frac{3}{2}}} \tag{3}$$

$$I_2=\frac{1}{\pi}\int_0^\pi \frac{\cos\theta d\theta}{(1-b\cos\theta)^{\frac{3}{2}}}, \tag{4}$$

where

$$b=\frac{2a\rho}{a^2+z^2+\rho^2}. \tag{5}$$

The integrals I_1 and I_2 can be expressed in terms of complete elliptic integrals if the following substitutions are made:

$$\phi=\tfrac{1}{2}(\pi-\theta) \tag{6}$$

$$k^2=2b/(1+b). \tag{7}$$

The results are

$$I_1=\frac{2}{\pi}\frac{(1-k^2/2)^{\frac{1}{2}}}{1-k^2}E \tag{8}$$

$$I_2=\frac{2}{\pi}\frac{(1-k^2/2)^{\frac{1}{2}}}{1-k^2}(2B-E), \tag{9}$$

where

$$E=\int_0^{\pi/2}(1-k^2\sin^2\phi)^{\frac{1}{2}}d\phi \tag{10}$$

and

$$B=\int_0^{\pi/2}\frac{\cos^2\phi d\phi}{(1-k^2\sin^2\phi)^{\frac{1}{2}}}. \tag{11}$$

SERIES EXPRESSIONS FOR INTEGRALS

If the integrands are expanded in series and integrated term by term, the following expressions are obtained for I_1 and I_2:

$$I_1=1+\frac{3\cdot5}{4\cdot4}b^2+\frac{3\cdot5\cdot7\cdot9}{4\cdot4\cdot8\cdot8}b^4+\frac{3\cdot5\cdot7\cdot9\cdot11\cdot13}{4\cdot4\cdot8\cdot8\cdot12\cdot12}b^6+\cdots \tag{12}$$

$$I_2=-\frac{3}{4}b+\frac{3\cdot5\cdot7}{4\cdot4\cdot8}b^3+\frac{3\cdot5\cdot7\cdot9\cdot11}{4\cdot4\cdot8\cdot8\cdot12}b^5+\cdots. \tag{13}$$

For values of b near unity these series converge slowly. With the aid of expressions for E and B in the neighborhood of $k=1$, I_1 and I_2 can be expressed as series in powers of $(1-b)$. Let

$$k'^2=1-k^2$$

Then[4]

$$E=1+\sum_{n=1}^{\infty}\frac{1^2\cdot3^2\cdots(2n-3)^2(2n-1)}{2^2\cdot4^2\cdots(2n-2)^2 2n}(\Lambda-e_n)k'^{2n} \tag{14}$$

$$B=1-\sum_{n=1}^{\infty}\frac{1^2\cdot3^2\cdots(2n-3)^2(2n-1)^2}{2^2\cdot4^2\cdots(2n-2)^2 2n}(\Lambda-b_n)k'^{2n}, \tag{15}$$

where

$$e_n=\frac{1}{2n}\left[1+\frac{6n-2}{1\cdot1\cdot3}+\frac{10n-4}{2\cdot3\cdot5}+\cdots+\frac{(2n-1)2n-(2n-2)}{(n-1)(2n-3)(2n-1)}\right] \tag{16}$$

$$b_n=\frac{1}{2n}\left[2n+1+\frac{2n-2}{1\cdot1\cdot3}+\frac{2n-4}{2\cdot3\cdot5}+\cdots+\frac{2}{(n-1)(2n-3)(2n-1)}\right] \tag{17}$$

$\Lambda=\ln 4/k'$.

The integrals I_1 and I_2 become

$$I_1=\frac{\sqrt{2}}{\pi}\left\{\left[\frac{1}{1-b}+\frac{1}{8}-\frac{13}{256}(1-b)-\frac{103}{1024}(1-b)^2-\frac{86567}{786432}(1-b)^3-\cdots\right]+\Lambda\left[\frac{1}{4}+\frac{15}{64}(1-b)+\frac{195}{1024}(1-b^2)+\frac{4935}{32768}(1-b)^3+\cdots\right]\right\} \tag{18}$$

$$I_2=\frac{\sqrt{2}}{\pi}\left\{\left[\frac{1}{1-b}+\frac{9}{8}+\frac{307}{256}(1-b)+\frac{1263}{1024}(1-b)^2+\frac{327219}{262144}(1-b)^3+\cdots\right]\right.$$

[4] F. Jahnke and E. Emde, *Tables of Functions* (Dover Publications, New York, 1933), pp. 73.

$$-\Lambda\left[\frac{3}{4}+\frac{57}{64}(1-b)+\frac{945}{1024}(1-b)^2+\frac{30345}{32768}(1-b)^3+\cdots\right]\Big\} \quad (19)$$

$\Lambda=\ln 4-\frac{1}{2}\ln[(1-b)/(1+b)]$.

THE TABLES

Table I lists values of I_1 for values of b in increments of 0.001 from 0 to 0.800. Table II lists corresponding values of I_2. Table III lists values of I_1, I_2, I_1-I_2, $(1-b)I_1$, and $(1-b)I_2$ for values of b in increments of 0.001 from 0.800 to 1.000. Table IV lists values of the same five functions for values of b in increments of 0.0001 from 0.9950 to 1.0000. The difference I_1-I_2 is included to facilitate the computation of H_s in the vicinity of $b=1$ (near the wire). In this case Eq. (1) involves the subtraction of two large quantities which are nearly equal. This difficulty can be circumvented by using I_1-I_2 as follows:

$$H_s=(H_0/s^3)\{(I_1-I_2)+[1-(\rho/a)]I_2\}.$$

The functions $(1-b)I_1$ and $(1-b)I_2$ are of value in computing I_1 and I_2 in the vicinity of $b=1$, since they lend themselves to linear interpolation.

TABLE I. Values of the integral I_1.

b	0	1	2	3	4	5	6	7	8	9
0.00	1.000 000	000 001	000 004	000 008	000 015	000 023	000 034	000 046	000 060	000 076
0.01	000 094	000 114	000 135	000 159	000 184	000 211	000 240	000 271	000 304	000 339
0.02	000 375	000 414	000 454	000 496	000 540	000 586	000 634	000 684	000 736	000 789
0.03	000 844	000 902	000 961	001 022	001 085	001 150	001 217	001 285	001 356	001 428
0.04	001 502	001 579	001 657	001 737	001 818	001 902	001 988	002 075	002 165	002 256
0.05	1.002 349	002 445	002 542	002 641	002 742	002 844	002 949	003 056	003 164	003 275
0.06	003 387	003 501	003 618	003 736	003 856	003 978	004 101	004 227	004 355	004 484
0.07	004 616	004 749	004 885	005 022	005 162	005 303	005 446	005 591	005 738	005 887
0.08	006 038	006 191	006 346	006 503	006 661	006 822	006 985	007 149	007 316	007 484
0.09	007 655	007 827	008 002	008 178	008 356	008 537	008 719	008 903	009 090	009 278
0.10	1.009 468	009 660	009 855	010 051	010 249	010 449	010 652	010 856	011 062	011 270
0.11	011 480	011 693	011 907	012 123	012 342	012 562	012 784	013 009	013 235	013 464
0.12	013 694	013 927	014 161	014 398	014 636	014 877	015 120	015 365	015 612	015 861
0.13	016 112	016 365	016 620	016 877	017 137	017 398	017 662	017 927	018 195	018 465
0.14	018 737	019 011	019 287	019 565	019 845	020 128	020 412	020 699	020 988	021 278
0.15	1.021 572	021 867	022 164	022 464	022 765	023 069	023 375	023 683	023 994	024 306
0.16	024 621	024 937	025 256	025 577	025 901	026 226	026 554	026 884	027 216	027 551
0.17	027 887	028 226	028 567	028 910	029 256	029 604	029 954	030 306	030 660	031 017
0.18	031 376	031 737	032 101	032 466	032 835	033 205	033 578	033 953	034 330	034 709
0.19	035 091	035 475	035 862	036 251	036 642	037 035	037 431	037 829	038 230	038 632
0.20	1.039 038	039 445	039 855	040 267	040 682	041 099	041 519	041 941	042 365	042 792
0.21	043 221	043 652	044 086	044 523	044 961	045 403	045 846	046 293	046 741	047 192
0.22	047 646	048 102	048 561	049 022	049 485	049 951	050 420	050 891	051 365	051 841
0.23	052 320	052 801	053 285	053 771	054 260	054 751	055 245	055 742	056 241	056 743
0.24	057 248	057 755	058 264	058 777	059 292	059 809	060 330	060 852	061 378	061 906
0.25	1.062 437	062 971	063 507	064 046	064 588	065 132	065 680	066 230	066 782	067 338
0.26	067 896	068 457	069 020	069 587	070 156	070 728	071 303	071 881	072 462	073 045
0.27	073 631	074 220	074 812	075 407	076 005	076 605	077 209	077 815	078 424	079 036
0.28	079 651	080 269	080 890	081 514	082 141	082 771	083 404	084 040	084 679	085 321
0.29	085 966	086 614	087 265	087 919	088 576	089 236	089 900	090 566	091 236	091 908
0.30	1.092 584	093 263	093 945	094 630	095 319	096 010	096 705	097 403	098 104	098 808
0.31	099 516	100 227	100 941	101 658	102 379	103 103	103 830	104 561	105 295	106 032
0.32	106 773	107 517	108 264	109 015	109 769	110 526	111 287	112 052	112 820	113 591
0.33	114 366	115 144	115 926	116 711	117 500	118 292	119 088	119 887	120 690	121 497
0.34	122 307	123 121	123 939	124 760	125 585	126 413	127 245	128 081	128 921	129 764
0.35	1.130 611	131 462	132 316	133 175	134 037	134 903	135 773	136 647	137 524	138 406
0.36	139 291	140 180	141 073	141 971	142 872	143 777	144 686	145 599	146 516	147 437
0.37	148 362	149 291	150 225	151 162	152 104	153 050	153 999	154 953	155 912	156 874
0.38	157 841	158 812	159 787	160 766	161 750	162 738	163 731	164 727	165 728	166 734
0.39	167 744	168 758	169 777	170 800	171 828	172 861	173 898	174 939	175 985	177 035
0.40	1.178 091	179 150	180 215	181 284	182 358	183 436	184 520	185 608	186 700	187 798
0.41	188 900	190 008	191 120	192 237	193 359	194 485	195 617	196 754	197 896	199 042
0.42	200 194	201 351	202 513	203 680	204 853	206 030	207 213	208 401	209 594	210 792
0.43	211 996	213 205	214 419	215 639	216 864	218 094	219 330	220 572	221 818	223 071
0.44	224 329	225 592	226 862	228 136	229 417	230 703	231 995	233 292	234 596	235 905

TABLE I.—*Continued.*

b	0	1	2	3	4	5	6	7	8	9
0.45	1.237 220	238 541	239 868	241 200	242 539	243 884	245 234	246 591	247 954	249 323
0.46	250 698	252 079	253 466	254 860	256 260	257 666	259 079	260 497	261 923	263 354
0.47	264 793	266 237	267 689	269 146	270 611	272 082	273 559	275 044	276 535	278 033
0.48	279 538	281 049	282 568	284 093	285 626	287 165	288 711	290 265	291 826	293 393
0.49	294 968	296 551	298 140	299 737	301 341	302 953	304 572	306 198	307 833	309 474
0.50	1.311 124	312 780	314 445	316 117	317 798	319 486	321 181	322 885	324 597	326 317
0.51	328 045	329 781	331 525	333 277	335 038	336 807	338 584	340 369	342 163	343 966
0.52	345 777	347 597	349 425	351 262	353 108	354 963	356 826	358 699	360 580	362 471
0.53	364 370	366 279	368 197	370 124	372 060	374 006	375 961	377 926	379 900	381 884
0.54	383 877	385 880	387 893	389 916	391 949	393 991	396 044	398 107	400 180	402 263
0.55	1.404 356	406 460	408 574	410 699	412 834	414 980	417 137	419 304	421 482	423 671
0.56	425 871	428 083	430 305	432 538	434 783	437 039	439 306	441 585	443 876	446 178
0.57	448 492	450 818	453 156	455 505	457 867	460 241	462 627	465 025	467 436	469 859
0.58	472 294	474 743	477 204	479 678	482 165	484 665	487 177	489 704	492 243	494 796
0.59	497 362	499 942	502 535	505 143	507 764	510 399	513 048	515 711	518 389	521 081
0.60	1.523 787	526 508	529 244	531 995	534 760	537 540	540 336	543 147	545 973	548 814
0.61	551 672	554 544	557 433	560 338	563 258	566 195	569 149	572 118	575 104	578 107
0.62	581 127	584 164	587 218	590 289	593 377	596 483	599 606	602 747	605 906	609 084
0.63	612 279	615 493	618 725	621 975	625 245	628 533	631 841	635 167	638 513	641 879
0.64	645 264	648 670	652 095	655 540	659 006	662 493	666 000	669 528	673 077	676 647
0.65	1.680 238	683 852	687 487	691 143	694 822	698 524	702 248	705 994	709 764	713 557
0.66	717 372	721 212	725 075	728 962	732 874	736 809	740 770	744 754	748 764	752 800
0.67	756 860	760 947	765 059	769 197	773 362	777 554	781 772	786 017	790 290	794 591
0.68	798 919	803 275	807 660	812 074	816 516	820 988	825 489	830 020	834 580	839 172
0.69	843 793	848 446	853 130	857 845	862 593	867 372	872 184	877 028	881 906	886 817
0.70	1.891 762	896 740	901 754	906 802	911 885	917 003	922 158	927 348	932 575	937 839
0.71	943 140	948 479	953 856	959 271	964 725	970 219	975 752	981 325	986 939	992 594
0.72	998 290	*004 027	*009 808	*015 630	*021 496	*027 406	*033 360	*039 358	*045 401	*051 490
0.73	2.057 625	063 807	070 036	076 312	082 636	089 009	095 432	101 904	108 427	115 000
0.74	121 625	128 302	135 032	141 815	148 652	155 544	162 491	169 494	176 553	183 670
0.75	2.190 844	198 077	205 370	212 722	220 135	227 609	235 146	242 745	250 409	258 136
0.76	265 930	273 789	281 715	289 709	297 772	305 904	314 107	322 381	330 728	339 147
0.77	347 641	356 209	364 854	373 576	382 376	391 255	400 214	409 255	418 378	427 584
0.78	436 875	446 252	455 715	465 267	474 908	484 639	494 462	504 379	514 390	524 496
0.79	534 700	545 002	555 404	565 908	576 514	587 225	598 041	608 965	619 998	631 141
0.80	2.642 397	653 767	665 252	676 855	688 577	700 420	712 386	724 476	736 694	749 040

TABLE II. Values of the integral I_2.

b	0	1	2	3	4	5	6	7	8	9
0.00	0.000 000	000 750	001 500	002 250	003 000	003 750	004 500	005 250	006 000	006 751
0.01	007 501	008 251	009 001	009 752	010 502	011 253	012 003	012 754	013 505	014 256
0.02	015 007	015 758	016 509	017 260	018 011	018 763	019 514	020 266	021 018	021 770
0.03	022 522	023 274	024 027	024 780	025 532	026 285	027 038	027 792	028 545	029 299
0.04	030 053	030 807	031 561	032 315	033 070	033 825	034 580	035 335	036 091	036 847
0.05	0.037 603	038 359	039 116	039 872	040 630	041 387	042 145	042 902	043 661	044 419
0.06	045 178	045 937	046 696	047 456	048 216	048 976	049 737	050 498	051 259	052 021
0.07	052 783	053 545	054 308	055 071	055 834	056 598	057 362	058 127	058 892	059 657
0.08	060 423	061 189	061 955	062 722	063 490	064 258	065 026	065 794	066 564	067 333
0.09	068 103	068 873	069 644	070 416	071 188	071 960	072 733	073 506	074 280	075 054
0.10	0.075 829	076 604	077 380	078 156	078 933	079 711	080 488	081 267	082 046	082 825
0.11	083 606	084 386	085 168	085 949	086 732	087 515	088 298	089 083	089 868	090 653
0.12	091 439	092 226	093 013	093 801	094 589	095 378	096 168	096 959	097 750	098 542
0.13	099 334	100 127	100 921	101 716	102 511	103 307	104 104	104 901	105 699	106 498
0.14	107 297	108 098	108 899	109 700	110 503	111 306	112 110	112 915	113 721	114 527
0.15	0.115 334	116 142	116 951	117 761	118 571	119 382	120 194	121 007	121 821	122 636
0.16	123 451	124 267	125 085	125 903	126 722	127 541	128 362	129 184	130 006	130 830
0.17	131 654	132 479	133 305	134 133	134 961	135 790	136 620	137 451	138 282	139 115
0.18	139 949	140 784	141 620	142 457	143 295	144 134	144 974	145 815	146 657	147 500
0.19	148 344	149 189	150 035	150 883	151 731	152 581	153 431	154 283	155 136	155 990

TABLE II.—*Continued.*

b	0	1	2	3	4	5	6	7	8	9
0.20	0.156 845	157 701	158 558	159 417	160 276	161 137	161 999	162 862	163 726	164 592
0.21	165 459	166 326	167 196	168 066	168 938	169 810	170 684	171 560	172 436	173 314
0.22	174 193	175 074	175 955	176 838	177 723	178 608	179 495	180 383	181 273	182 164
0.23	183 056	183 950	184 845	185 741	186 639	187 538	188 439	189 341	190 244	191 149
0.24	192 055	192 963	193 872	194 783	195 695	196 609	197 524	198 440	199 359	200 278
0.25	0.201 199	202 122	203 046	203 972	204 900	205 828	206 759	207 691	208 625	209 560
0.26	210 497	211 436	212 376	213 318	214 261	215 206	216 153	217 101	218 052	219 004
0.27	219 957	220 912	221 869	222 828	223 789	224 751	225 715	226 681	227 649	228 618
0.28	229 589	230 562	231 537	232 514	233 493	234 473	235 455	236 440	237 426	238 414
0.29	239 404	240 396	241 389	242 385	243 383	244 383	245 384	246 388	247 393	248 401
0.30	0.249 411	250 423	251 436	252 452	253 470	254 490	255 512	256 536	257 563	258 591
0.31	259 622	260 654	261 689	262 726	263 765	264 807	265 851	266 897	267 945	268 995
0.32	270 048	271 103	272 160	273 219	274 281	275 345	276 412	277 481	278 552	279 626
0.33	280 702	281 780	282 861	283 944	285 030	286 118	287 208	288 302	289 397	290 495
0.34	291 596	292 699	293 805	294 913	296 024	297 138	298 254	299 372	300 494	301 618
0.35	0.302 745	303 874	305 006	306 141	307 279	308 419	309 562	310 708	311 856	313 008
0.36	314 162	315 319	316 479	317 642	318 808	319 976	321 148	322 322	323 500	324 680
0.37	325 864	327 050	328 240	329 432	330 628	331 826	333 028	334 233	335 440	336 652
0.38	337 866	339 083	340 304	341 527	342 754	343 985	345 218	346 455	347 695	348 939
0.39	350 185	351 435	352 689	353 946	355 206	356 470	357 737	359 008	360 282	361 560
0.40	0.362 841	364 126	365 414	366 706	368 002	369 301	370 604	371 911	373 221	374 535
0.41	375 852	377 174	378 499	379 828	381 161	382 498	383 839	385 183	386 532	387 884
0.42	389 240	390 601	391 965	393 334	394 706	396 083	397 463	398 848	400 237	401 630
0.43	403 028	404 429	405 835	407 245	408 659	410 078	411 501	412 928	414 360	415 797
0.44	417 237	418 682	420 132	421 586	423 045	424 509	425 977	427 449	428 927	430 409
0.45	0.431 896	433 387	434 883	436 385	437 891	439 401	440 917	442 438	443 963	445 494
0.46	447 030	448 570	450 116	451 667	453 223	454 784	456 351	457 922	459 499	461 082
0.47	462 669	464 262	465 860	467 464	469 074	470 688	472 308	473 934	475 566	477 203
0.48	478 846	480 494	482 148	483 808	485 474	487 146	488 824	490 507	492 197	493 892
0.49	495 594	497 301	499 015	500 735	502 461	504 193	505 932	507 677	509 428	511 186
0.50	0.512 950	514 720	516 498	518 281	520 071	521 868	523 672	525 482	527 299	529 123
0.51	530 954	532 792	534 637	536 488	538 347	540 213	542 086	543 966	545 853	547 748
0.52	549 650	551 559	553 476	555 400	557 332	559 271	561 218	563 173	565 135	567 106
0.53	569 084	571 070	573 064	575 065	577 075	579 093	581 119	583 154	585 197	587 248
0.54	589 307	591 375	593 451	595 536	597 629	599 732	601 842	603 962	606 091	608 228
0.55	0.610 375	612 530	614 695	616 869	619 052	621 244	623 446	625 657	627 878	630 108
0.56	632 348	634 598	636 857	639 126	641 405	643 694	645 994	648 303	650 622	652 952
0.57	655 292	657 643	660 004	662 376	664 758	667 152	669 556	671 970	674 396	676 833
0.58	679 281	681 740	684 211	686 693	689 187	691 692	694 208	696 737	699 277	701 829
0.59	704 393	706 970	709 558	712 159	714 772	717 398	720 036	722 687	725 351	728 027
0.60	0.730 717	733 419	736 135	738 864	741 607	744 362	747 132	749 915	752 712	755 523
0.61	758 348	761 187	764 040	766 908	769 790	772 687	775 598	778 525	781 466	784 422
0.62	787 394	790 381	793 383	796 401	799 434	802 484	805 549	808 631	811 728	814 842
0.63	817 973	821 120	824 284	827 465	830 663	833 878	837 111	840 361	843 628	846 914
0.64	850 217	853 538	856 878	860 236	863 613	867 009	870 423	873 856	877 309	880 781
0.65	0.884 273	887 784	891 315	894 867	898 438	902 031	905 643	909 277	912 931	916 607
0.66	920 304	924 023	927 764	931 526	935 311	939 118	942 948	946 800	950 676	954 574
0.67	958 496	962 442	966 412	970 406	974 424	978 466	982 534	986 626	990 744	994 887
0.68	999 056	*003 251	*007 473	*011 720	*015 995	*020 296	*024 625	*028 982	*033 366	*037 778
0.69	1.042 219	046 688	051 186	055 714	060 270	064 857	069 474	074 121	078 799	083 508
0.70	1.088 249	093 021	097 825	102 661	107 531	112 433	117 368	122 337	127 340	132 378
0.71	137 450	142 558	147 701	152 880	158 095	163 346	168 635	173 962	179 326	184 728
0.72	190 169	195 649	201 169	206 729	212 329	217 970	223 652	229 376	235 142	240 951
0.73	246 804	252 699	258 640	264 624	270 654	276 730	282 852	289 021	295 237	301 501
0.74	307 813	314 175	320 586	327 047	333 559	340 123	346 738	353 406	360 128	366 903
0.75	1.373 733	380 618	387 559	394 556	401 611	408 724	415 896	423 127	430 418	437 770
0.76	445 184	452 661	460 201	467 805	475 474	483 209	491 010	498 879	506 817	514 824
0.77	522 901	531 049	539 270	547 563	555 931	564 373	572 892	581 488	590 162	598 916
0.78	607 749	616 664	625 662	634 744	643 910	653 163	662 503	671 932	681 450	691 060
0.79	700 762	710 558	720 449	730 437	740 523	750 708	760 994	771 383	781 876	792 474
0.80	1.803 179	813 994	824 918	835 955	847 106	858 372	869 756	881 259	892 884	904 631

TABLE III.

b	I_1	I_2	I_1-I_2	$(1-b)I_1$	$(1-b)I_2$
0.800	2.642 397	1.803 179	0.839 218	0.528 479	0.360 636
0.801	653 767	813 994	839 773	528 100	360 985
0.802	665 252	824 918	840 334	527 720	361 334
0.803	676 855	835 955	840 900	527 340	361 683
0.804	688 577	847 106	841 471	526 961	362 033
0.805	2.700 420	1.858 372	0.842 048	0.526 582	0.362 383
0.806	712 386	869 756	842 629	526 203	362 733
0.807	724 476	881 259	843 217	525 824	363 083
0.808	736 694	892 884	843 810	525 445	363 434
0.809	749 040	904 631	844 408	525 067	363 785
0.810	2.761 517	1.916 504	0.845 012	0.524 688	0.364 136
0.811	744 126	928 504	845 622	524 310	364 487
0.812	786 871	940 633	846 238	523 932	364 839
0.813	799 752	952 893	846 859	523 554	365 191
0.814	812 773	965 287	847 486	523 176	365 543
0.815	2.825 936	1.977 816	0.848 119	0.522 798	0.365 896
0.816	839 242	1.990 483	848 758	522 420	366 249
0.817	852 694	2.003 290	849 404	522 043	366 602
0.818	866 295	016 240	850 055	521 666	366 956
0.819	880 047	029 334	850 713	521 288	367 310
0.820	2.893 952	2.042 576	0.851 376	0.520 911	0.367 664
0.821	908 014	055 967	852 047	520 534	368 018
0.822	922 234	069 511	852 723	520 158	368 373
0.823	936 616	083 210	853 406	519 781	368 728
0.824	951 161	097 066	854 096	519 404	369 083
0.825	2.965 874	2.111 082	0.854 792	0.519 028	0.369 439
0.826	980 756	125 262	855 495	518 652	369 796
0.827	2.995 812	139 607	856 204	518 275	370 152
0.828	3.011 043	154 121	856 921	517 899	370 509
0.829	026 452	168 808	857 645	517 523	370 866
0.830	3.042 044	2.183 669	0.858 375	0.517 147	0.371 224
0.831	057 821	198 708	859 113	516 772	371 582
0.832	073 786	213 928	859 858	516 396	371 940
0.833	089 943	229 333	860 610	516 020	372 299
0.834	106 295	244 926	861 369	515 645	372 658
0.835	3.122 846	2.260 710	0.862 136	0.515 270	0.373 017
0.836	139 600	276 689	862 911	514 894	373 377
0.837	156 559	292 867	863 693	514 519	373 737
0.838	173 729	309 246	864 483	514 144	374 098
0.839	191 112	325 832	865 280	513 769	374 459
0.840	3.208 714	2.342 628	0.866 086	0.513 394	0.374 820
0.841	226 537	359 638	866 899	513 019	375 182
0.842	244 586	376 865	867 721	512 645	375 545
0.843	262 866	394 315	868 551	512 270	375 907
0.844	281 381	411 991	869 390	511 895	376 271
0.845	3.300 135	2.429 898	0.870 237	0.511 521	0.376 634
0.846	319 133	448 041	871 092	511 147	376 998
0.847	338 380	466 424	871 956	510 772	377 363
0.848	357 881	485 052	872 829	510 398	377 728
0.849	377 640	503 929	873 711	510 024	378 093
0.850	3.397 664	2.523 062	0.874 602	0.509 650	0.378 459
0.851	417 956	542 454	875 502	509 275	378 826
0.852	438 523	562 112	876 411	508 901	379 193
0.853	459 370	582 041	877 330	508 527	379 560
0.854	480 504	602 246	878 258	508 154	379 928
0.855	3.501 929	2.622 733	0.879 196	0.507 780	0.380 296
0.856	523 652	643 508	880 143	507 406	380 665
0.857	545 679	664 578	881 101	507 032	381 035
0.858	568 016	685 948	882 068	506 658	381 405
0.859	590 671	707 625	883 046	506 285	381 775
0.860	3.613 650	2.729 615	0.884 035	0.505 911	0.382 146
0.861	636 959	751 926	885 033	505 537	382 518
0.862	660 607	774 564	886 043	505 164	382 890
0.863	684 600	797 537	887 063	504 790	383 263
0.864	708 946	820 851	888 094	504 417	383 636
0.865	3.733 653	2.844 516	0.889 137	0.504 043	0.384 010
0.866	758 729	868 538	890 191	503 670	384 384
0.867	784 182	892 925	891 256	503 296	384 759
0.868	810 020	917 687	892 333	502 923	385 135
0.869	836 254	942 832	893 422	502 549	385 511
0.870	3.862 891	2.968 368	0.894 523	0.502 176	0.385 888
0.871	889 940	2.994 304	895 636	501 802	386 265
0.872	917 413	3.020 651	896 762	501 429	386 643
0.873	945 318	047 418	897 900	501 055	387 022
0.874	3.973 666	074 615	899 051	500 682	387 402
0.875	4.002 468	3.102 253	0.900 215	0.500 308	0.387 782
0.876	031 734	130 341	901 393	499 935	388 162
0.877	061 476	158 892	902 583	499 562	388 544
0.878	091 705	187 917	903 788	499 188	388 926
0.879	122 434	217 427	905 007	498 814	389 309

TABLE III.—*Continued.*

b	I_1	I_2	I_1-I_2	$(1-b)I_1$	$(1-b)I_2$
0.880	4.153 674	3.247 435	0.906 239	0.498 441	0.389 692
0.881	185 439	277 953	907 486	498 067	390 076
0.882	217 743	308 995	908 748	497 694	390 461
0.883	250 598	340 573	910 025	497 320	390 847
0.884	284 019	372 703	911 317	496 946	391 234
0.885	4.318 022	3.405 398	0.912 624	0.496 572	0.391 621
0.886	352 620	438 673	913 947	496 199	392 009
0.887	387 830	472 544	915 286	495 825	392 397
0.888	423 668	507 027	916 641	495 451	392 787
0.889	460 152	542 139	918 013	495 077	393 177
0.890	4.497 298	3.577 896	0.919 402	0.494 703	0.393 569
0.891	535 125	614 317	920 808	494 329	393 961
0.892	573 652	651 421	922 232	493 954	394 353
0.893	612 899	689 226	923 673	493 580	394 747
0.894	652 885	727 752	925 133	493 206	395 142
0.895	4.693 631	3.767 020	0.926 611	0.492 831	0.395 537
0.896	735 161	807 053	928 108	492 457	395 933
0.897	777 496	847 871	929 624	492 082	396 331
0.898	820 659	889 499	931 161	491 707	396 729
0.899	864 677	931 960	932 717	491 332	397 128
0.900	4.909 574	3.975 280	0.934 294	0.490 957	0.397 528
0.901	4.955 376	4.019 485	935 891	490 582	397 929
0.902	5.002 112	064 602	937 510	490 207	398 331
0.903	049 810	110 659	939 151	489 832	398 734
0.904	098 501	157 687	940 814	489 456	399 138
0.905	5.148 215	4.205 715	0.942 500	0.489 080	0.399 543
0.906	198 985	254 776	944 209	488 705	399 949
0.907	250 845	304 903	945 942	488 329	400 356
0.908	303 831	356 132	947 699	487 952	400 764
0.909	357 980	408 499	949 481	487 576	401 173
0.910	5.413 330	4.462 042	0.951 288	0.487 200	0.401 584
0.911	469 922	516 801	953 121	486 823	401 995
0.912	527 798	572 817	954 981	486 446	402 408
0.913	587 002	630 134	956 868	486 069	402 822
0.914	647 580	688 798	958 782	485 692	403 237
0.915	5.709 582	4.748 856	0.960 726	0.485 314	0.403 653
0.916	773 057	810 359	962 698	484 937	404 070
0.917	838 059	873 359	964 700	484 559	404 489
0.918	904 644	4.937 911	966 733	484 181	404 909
0.919	5.972 869	5.004 072	968 797	483 802	405 330
0.920	6.042 798	5.071 904	0.970 894	0.483 424	0.405 752
0.921	114 493	141 470	973 023	483 045	406 176
0.922	188 024	212 837	975 187	482 666	406 601
0.923	263 460	286 076	977 385	482 286	407 028
0.924	340 879	361 260	979 618	481 907	407 456
0.925	6.420 358	5.438 469	0.981 889	0.481 527	0.407 885
0.926	501 981	517 783	984 197	481 147	408 316
0.927	585 836	599 292	986 544	480 766	408 748
0.928	672 015	683 085	988 931	480 385	409 182
0.929	760 618	769 260	991 358	480 004	409 617
0.930	6.851 748	5.857 920	0.993 828	0.479 622	0.410 054
0.931	6.945 514	5.949 173	996 341	479 240	410 493
0.932	7.042 032	6.043 133	0.998 899	478 858	410 933
0.933	141 426	139 923	1.001 503	478 476	411 375
0.934	243 827	239 672	004 154	478 093	411 818
0.935	7.349 371	6.342 517	1.006 854	0.477 709	0.412 264
0.936	458 208	448 603	009 605	477 325	412 711
0.937	570 494	558 086	012 408	476 941	413 159
0.938	686 394	671 130	015 264	476 556	413 610
0.939	806 087	787 911	018 176	476 171	414 063
0.940	7.929 762	6.908 617	1.021 146	0.475 786	0.414 517
0.941	8.057 622	7.033 447	024 175	475 400	414 973
0.942	189 882	162 617	027 265	475 013	415 432
0.943	326 774	296 355	030 419	474 626	415 892
0.944	468 546	434 906	033 639	474 239	416 355
0.945	8.615 463	7.578 535	1.036 928	0.473 850	0.416 819
0.946	767 811	727 524	040 287	473 462	417 286
0.947	8.925 898	7.882 177	043 720	473 073	417 755
0.948	9.090 053	8.042 823	047 230	472 683	418 227
0.949	260 634	209 815	050 819	472 292	418 701
0.950	9.438 025	8.383 534	1.054 491	0.471 901	0.419 177
0.951	9.622 644	8.564 395	058 249	471 510	419 655
0.952	9.814 941	8.752 844	062 097	471 117	420 137
0.953	10.015 406	8.949 368	066 038	470 724	420 620
0.954	10.224 571	9.154 494	070 077	470 330	421 107
0.955	10.443 017	9.368 798	1.074 219	0.469 936	0.421 596
0.956	10.671 373	9.592 907	078 467	469 540	422 088
0.957	10.910 333	9.827 507	083 826	469 144	422 583
0.958	11.160 652	10.073 349	087 303	468 747	423 081
0.959	11.423 161	10.331 258	091 903	468 350	423 582

TABLE III.—*Continued.*

b	I_1	I_2	I_1-I_2	$(1-b)I_1$	$(1-b)I_2$
0.960	11.698 773	10.602 141	1.096 632	0.467 951	0.424 086
0.961	11.988 496	10.887 000	101 496	467 551	424 593
0.962	12.293 442	11.186 938	106 503	467 151	425 104
0.963	12.614 845	11.503 184	111 661	466 749	425 618
0.964	12.954 075	11.837 097	116 977	466 347	426 136
0.965	13.312 659	12.190 197	1.122 462	0.465 943	0.426 657
0.966	13.692 304	12.564 179	128 125	465 538	427 182
0.967	14.094 922	12.960 947	133 976	465 132	427 711
0.968	14.522 667	13.382 640	140 027	464 725	428 244
0.969	14.977 968	13.831 675	146 292	464 317	428 782
0.970	15.463 578	14.310 793	1.152 785	0.463 907	0.429 324
0.971	15.982 632	14.823 111	159 520	463 496	429 870
0.972	16.538 710	15.372 193	166 516	463 084	430 421
0.973	17.135 923	15.962 131	173 792	462 670	430 978
0.974	17.779 016	16.597 647	181 369	462 254	431 539
0.975	18.473 491	17.284 220	1.189 270	0.461 837	0.432 106
0.976	19.225 766	18.028 243	197 523	461 418	432 678
0.977	20.043 378	18.837 219	206 159	460 998	433 256
0.978	20.935 231	19.720 020	215 210	460 575	433 840
0.979	21.911 926	20.687 208	224 717	460 150	434 431
0.980	22.986 183	21.751 458	1.234 725	0.459 724	0.435 029
0.981	24.173 400	22.928 116	245 285	459 295	435 634
0.982	25.492 397	24.235 940	256 457	458 863	436 247
0.983	26.966 418	25.698 105	268 313	458 429	436 868
0.984	28.624 519	27.343 584	280 935	457 992	437 497
0.985	30.503 504	29.209 079	1.294 425	0.457 553	0.438 136
0.986	32.650 687	31.341 786	308 901	457 110	438 785
0.987	35.127 941	33.803 428	324 513	456 663	439 445
0.988	38.017 758	36.676 316	341 442	456 213	440 116
0.989	41.432 623	40.072 702	359 921	455 759	440 800
0.990	45.530 008	44.149 762	1.380 246	0.455 300	0.441 498
0.991	50.537 365	49.134 555	402 810	454 836	442 211
0.992	56.795 854	55.367 710	428 144	454 367	442 942
0.993	64.841 565	63.384 573	456 992	453 891	443 692
0.994	75.567 942	74.077 495	490 447	453 408	444 465
0.995	90.583 117	89.052 914	1.530 203	0.452 916	0.445 265
0.996	113.103 214	111.524 117	579 098	452 413	446 096
0.997	150.632 202	148.989 745	642 457	451 897	446 969
0.998	225.680 980	223.948 737	732 243	451 362	447 897
0.999	450.798 638	448.911 987	886 651	450 799	448 912
1.000	∞	∞	∞	0.450 158	0.450 158

TABLE IV.

b	I_1	I_2	I_1-I_2	$(1-b)I_1$	$(1-b)I_2$
0.9950	90.583 12	89.052 91	1.530 203	0.452 9156	0.445 2646
1	92.421 60	90.886 98	534 619	452 8658	445 3462
2	94.336 66	92.797 53	539 129	452 8160	445 4281
3	96.333 19	94.789 45	543 736	452 7660	445 5104
4	98.416 50	96.868 06	548 445	452 7159	445 5931
0.9955	100.592 38	99.039 12	1.553 259	0.452 6657	0.445 6760
6	102.867 13	101.308 95	558 184	452 6154	445 7594
7	105.247 66	103.684 44	563 225	452 5649	445 8431
8	107.741 52	106.173 13	568 386	452 5144	445 9272
9	110.357 00	108.783 32	573 675	452 4637	446 0116
0.9960	113.103 21	111.524 12	1.579 098	0.452 4129	0.446 0965
1	115.990 23	114.405 57	584 660	452 3619	446 1817
2	119.029 16	117.438 79	590 370	452 3108	446 2674
3	122.232 32	120.636 08	596 235	452 2596	446 3535
4	125.613 39	124.011 12	602 264	452 2082	446 4400
0.9965	129.187 62	127.579 15	1.608 466	0.452 1567	0.446 5270
6	132.972 05	131.357 20	614 851	452 1050	446 6145
7	136.985 80	135.364 37	621 430	452 0531	446 7024
8	141.250 35	139.622 14	628 215	452 0011	446 7908
9	145.789 99	144.154 77	635 219	451 9490	446 8798
0.9970	150.632 20	148.989 75	1.642 457	0.451 8966	0.446 9692
1	155.808 30	154.158 36	649 944	451 8441	447 0592
2	161.354 06	159.696 36	657 698	451 7914	447 1498
3	167.310 54	165.644 80	665 739	451 7385	447 2410
4	173.725 13	172.051 04	674 088	451 6853	447 3327
0.9975	180.652 81	178.970 04	1.682 769	0.451 6320	0.447 4251
6	188.157 70	186.465 89	691 809	451 5785	447 5181
7	196.315 10	194.613 86	701 239	451 5247	447 6119
8	205.213 96	203.502 87	711 094	451 4707	447 7063
9	214.960 22	213.238 81	721 414	451 4165	447 8015
0.9980	225.680 98	223.948 74	1.732 243	0.451 3620	0.447 8975
1	237.530 09	235.786 46	743 635	451 3072	447 9943
2	250.695 61	248.939 96	755 649	451 2521	448 0919
3	265.409 84	263.641 48	768 358	451 1967	448 1905
4	281.963 14	280.181 29	781 846	451 1410	448 2901
0.9985	300.723 32	298.927 11	1.796 213	0.451 0850	0.448 3907
6	322.163 26	320.351 68	811 581	451 0286	448 4924
7	346.901 35	345.073 26	828 099	450 9718	448 5952
8	375.762 11	373.916 16	845 950	450 9145	448 6994
9	409.869 85	408.004 48	865 368	450 8568	448 8049
0.9990	450.798 6	448.912 0	1.886 651	0.450 7986	0.448 9120
1	500.822 1	498.911 9	910 194	450 7399	449 0207
2	563.350 6	561.414 1	936 530	450 6805	449 1313
3	643.743 5	641.777 1	1.966 408	450 6204	449 2440
4	750.932 6	748.931 7	2.000 923	450 5596	449 3590
0.9995	900.995 6	898.953 8	2.041 775	0.450 4978	0.449 4769
6	1126.087 2	1123.995 3	091 809	450 4349	449 5981
7	1501.235 1	1499.078 7	156 364	450 3705	449 7236
8	2251.521 5	2249.274 0	247 420	450 3043	449 8548
9	4502.351 2	4499.948 0	403 215	450 2351	449 9948
1.0000	∞	∞	∞	0.450 1582	0.450 1582

Effect of Coating Composition of Oxide-Coated Cathodes on Electron Emission

E. G. Widell and R. A. Hellar
Tube Department, RCA Victor Division, Harrison, New Jersey
(Received June 6, 1950)

Investigation of oxide-coated cathodes has been made to determine the effect on electron emission of varying the proportions of emitting oxides. The results show that maximum electron emission under saturation conditions is obtained from a solid solution containing strontium oxide (SrO) and barium oxide (BaO) in an approximate molecular ratio of 7 to 3.

These results also indicate that maximum size of the co-precipitated alkaline-earth carbonate particle occurs at the same molecular ratio as that giving maximum electron emission from the oxides.

When the saturation emission obtained with a square-wave voltage pulse was measured, the current pulse as observed on a synchroscope was also a square wave and showed no measurable decay characteristic for a pulse duration time of 10 microseconds.

INTRODUCTION

MANY investigations have been made of the thermionic emission from alkaline-earth oxide-coated cathodes. However, only a few have considered the relation between emission capability and coating composition. In general, coatings have followed a 50–50 ratio of barium and strontium oxides either as weight percent or as molecular percent and have been applied in the form of a mechanical mixture or co-precipitate of the carbonates.

Arnold[1] in 1920 reported that the highest emission was obtained from BaO alone but an equal quantity of SrO was added to improve life. Simon[2] in 1927 found two emission maxima, the first at about 85 percent SrO–15 percent BaO, and the second at about 85 percent BaO–15 percent SrO (weight percent). BaO alone yielded about 70 percent of the emission obtained from a 50–50 ratio. No other investigator finds two maxima. In 1933, Benjamin and Rooksby[3] reported maximum electron emission from a molecular weight ratio of 63 percent SrO and 37 percent BaO. Huber[4] in 1941 obtained a maximum for the BaO—SrO system at about 55 percent BaO with an emission value about ten times as great as from BaO alone. Veenemans[5] in 1943 published a curve showing maximum emission at approximately 75 molecular percent of SrO. Weinreich[6] reported in 1945 that a BaO—SrO cathode coating containing 98.2 percent SrO yielded approximately the same surface emission as coatings containing lesser amounts of SrO. Eisenstein[7] in 1946 stated that Fineman of the M.I.T. Radiation Laboratory obtained the same pulsed electron emission from BaO alone as from a coating containing 50 percent BaO–50 percent SrO.

In view of all these apparent contradictions and because present-day requirements for high electron-emission currents under pulsed conditions again focus attention on types of oxide coatings which yield maximum currents, it was thought desirable to reinvestigate this subject.

MATERIALS AND CONSTRUCTION

The carbonates were prepared by co-precipitation from nitrate solutions using sodium carbonate as the precipitating agent with constant conditions of precipitation except for change in ratios of the two components. The strontium carbonate actually contained 1.0 percent $BaCO_3$ while the barium carbonate contained 0.2 percent $SrCO_3$, but for purposes of comparison the values are listed as 100 percent for both carbonates and oxides.

The cathode sleeves were made of Grade *A* nickel which, by spectrographic analysis, contained 0.2 percent silicon, 0.045 percent magnesium, 0.10 percent manganese, 0.07 percent iron, 0.04 percent copper, and less than 0.02 percent titanium. Oxide inclusions were not measured. The sleeves were 25 millimeters long and had an outside diameter of 0.762 millimeters. They were cleaned by electropolishing before being coated in order to remove surface contaminations.

The carbonates were milled for 45 hours under identical conditions in the usual organic solvents with nitrocellulose as a binder, and were sprayed on the central 10-millimeter portion of the cathode sleeves. The average weight of the sprayed coating was 3.4 mg/cm² with a thickness of 0.044 millimeter giving an average coating density of about 0.73 g/cm³.

At least five tubes were used for testing each ratio of oxides in order to obtain more dependable results than could be obtained by the use of a single tube. In order to obtain even temperatures throughout the length of the coated cathodes, only 10 millimeters of the central portions of each cathode sleeve were coated. Optical pyrometer measurements showed a variation of less than 10°K between the center and the ends of the coated portion. The bare ends of the sleeves were carefully cleaned to prevent unwanted electron emission. Tests with a specially constructed tube showed that the coated portion gave a thousand times more emission than the bare portions; therefore, emission from the

[1] H. D. Arnold, Phys. Rev. 16, 70 (1920).
[2] H. Simon, Zeits. f. techn. Phys. 8, 434 (1927).
[3] M. Benjamin and H. P. Rooksby, Phil. Mag. 15, 810 (1933).
[4] H. Huber, dissertation, Berlin, 1941.
[5] C. P. Veenemans, Nederl. Tijdschr. Natuurkunde 10, 1 (1943).
[6] M. O. Weinreich, Rev. Gén. Elec. (August, 1945).
[7] A. Eisenstein, J. App. Phys. 17, 654 (1946).

TABLE I

Carbonates					
Molecular % $BaCO_3$	Molecular % $SrCO_3$	"Average" particle size, microns	Bulk density, g/cm³	Sprayed cathode-coating density, g/cm³	Cathode brightness temperature, °K
0	100	1 32	0 46	0 76	998
10	90	1 87	0 39	0 85	1010
20	80	3 08	0 35	0 73	1000
30	70	3 35	0 33	0 67	996
40	60	2 92	0 37	0 73	998
50	50	2 83	0 40	0 73	1006
60	40	2 59	0 42	0 74	989
70	30	2 11	—	0 74	991
80	20	1 96	0 48	0 71	992
100	0	1 23	0 63	0 79	975

bare portions could be neglected The cathodes were mounted in a standard diode in which carbonized plates with inside diameters of 1.78 millimeters were used. Each tube received the same exhaust on an all-glass system using liquid air with mercury diffusion pumps and was sealed off after flashing a Ba—Mg getter. All tubes were given the same aging schedule

MEASUREMENTS

Structural Characteristics of the Carbonates

The particle sizes of the carbonates as precipitated were measured by the Andreasen pipette sedimentation method. The "average" particle size reported is the particle size for the 50 percent point on the curve showing cumulative weight percent *versus* particle size. These data are shown in Table I and are compared with emission data in Fig. 6. X-ray diffraction patterns indicate that the carbonate crystals of the 50–50 ratio have an aragonite structure. Electron microscope pictures of 100 percent $BaCO_3$, 70 percent $BaCO_3$–30 percent $SrCO_3$, 30 percent $BaCO_3$–70 percent $SrCO_3$, and 100 percent $SrCO_3$ are shown in Fig. 1. These pictures show the change in particle size with change in ratios of the carbonates. Particle or crystallite measurements of the 70 percent $BaCO_3$–30 percent $SrCO_3$ were made by the method of low angle scattering of x-rays using $K\alpha$-radiation at 35 kilovolts and 15 milliamperes, filtered by nickel; the x-ray film was spaced 50 centimeters from the specimens. These results are shown in Table II. The particle size as determined by this method may be the measurements of single or small groups of crystallites since the diameters are less than 200 angstroms, whereas the particle sizes as measured by the electron microscope or by the Andreasen pipette are a hundred times greater The above data show that both the particles and the crystallites of the 70 percent $SrCO_3$–30 percent $BaCO_3$ are about 1 5 times larger than those of the 30 percent $SrCO_3$–70 percent $BaCO_3$

Cathode Temperatures

Cathode temperature measurements were made with a disappearing-filament type of optical pyrometer ($\lambda = 0.665\mu$). The cathodes were initially adjusted to a brightness temperature of about 1000°K in order to

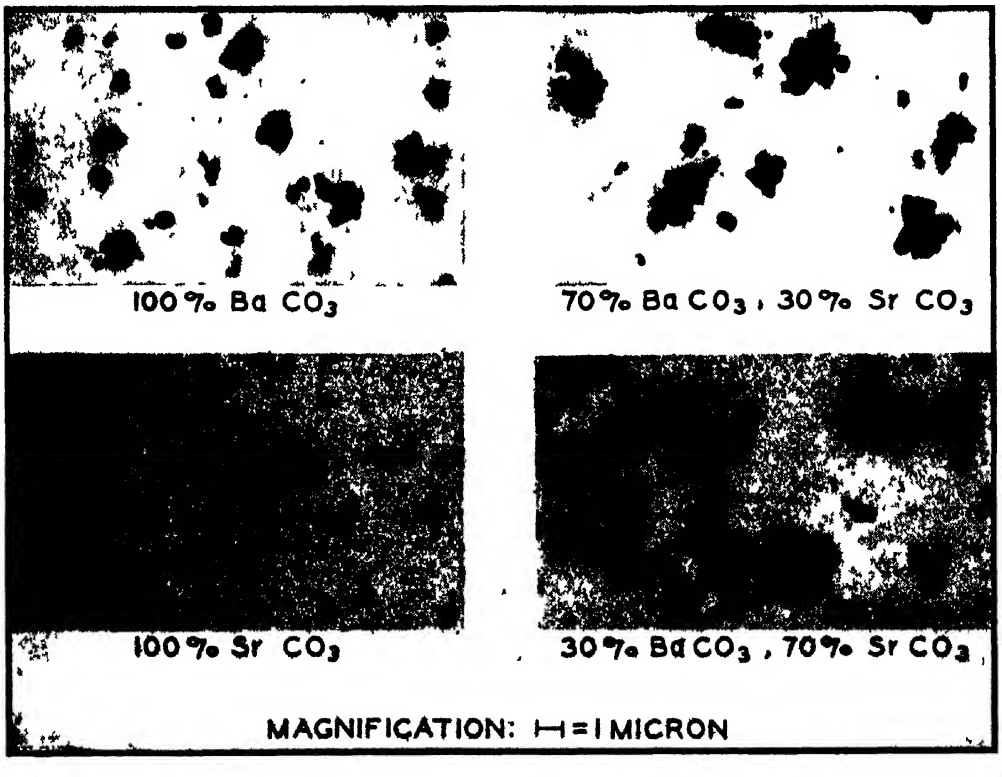


FIG 1 Electron micrographs of carbonate particles.

TABLE II. Particle size measurement by low angle scattering of x-rays.*

Type of carbonate	$r=1.0$	1.5	2.0	2.5	3.0
70% $SrCO_3$–30% $BaCO_3$	$R=105$	79	67	48	38
30% $SrCO_3$–70% $BaCO_3$	$R=63$	58	44	34	32

* r = distance from center of beam on film in millimeters. R = radius of particle in angstrom units.

limit electron emission and thus limit the heating up of the tube parts and cathode coatings during testing. The true temperature was then about 1030°K based on a spectral-emissivity (0.665μ) value of 0.56 for this type of coating obtained from unpublished data in this laboratory. The cathode temperatures shown in Table I are brightness temperatures (with no correction for glass absorption) and are averages for five tubes in each test.

Electron Emission

The emission measurements were made on a pulse emission tester designed to produce microsecond rectangular pulses. The pulse duration was set at 3 microseconds and the repetition rate at 500 pulses per second. A synchroscope showed wave forms of voltage and current to be square. Each measurement plotted represents an averaged reading for five or more tubes. The results are shown in Figs. 2, 3, and 4. Figure 5 compares these data with those of Benjamin and Rooksby, and of Veenemans.

Emission Decay

For measurements of relative values of emission for the various ratios of oxides under pulse-testing conditions, it may be necessary to consider the effect of emission decay characteristics as reported by Sproull[8] who found a considerable decay of emission during a 10-microsecond pulse. Coomes[9] also mentions this effect, stating that Ramsay of the Bartol Foundation found

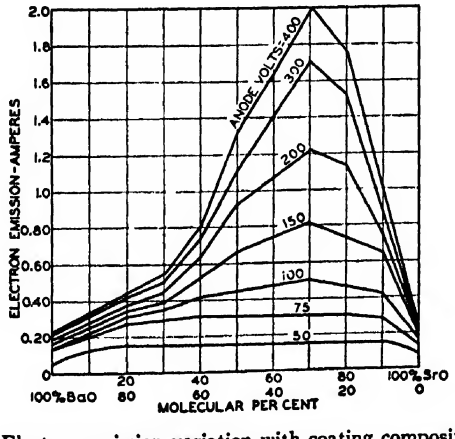


FIG. 2. Electron emission variation with coating compositions at different anode potentials.

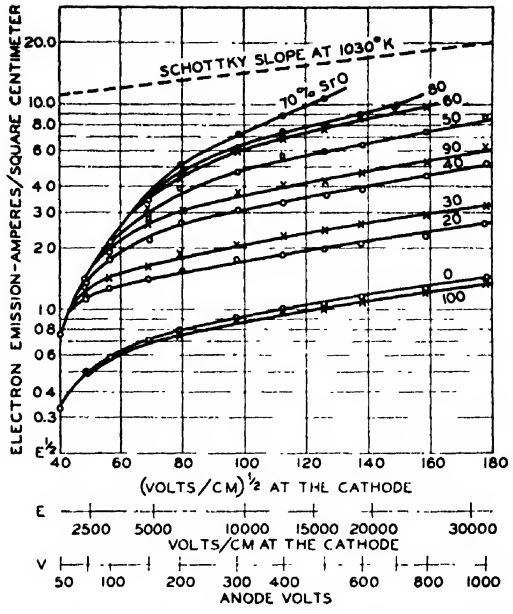


FIG. 3. Electron emission for various coating compositions. Schottky plot. $\mathrm{Log}_{10}(I/\mathrm{cm}^2)$ *versus* $E^{\frac{1}{2}}$, E, V.

short-time (10^{-6} second) emission decay in aged cathodes but not in very active cathodes. Observations for this effect were made on the cathodes with a 10-microsecond pulse. The emission pulse, observed on a synchroscope, had a square wave form and showed no measurable decay during this pulse period.

Decay effects have been observed in this laboratory from cathodes in tubes of the same construction but made for testing different materials. The decay effect in this case, about 40 percent decay during a 10-microsecond pulse, occurred only in those tubes which showed a heavy cathode poisoning effect as evidenced by

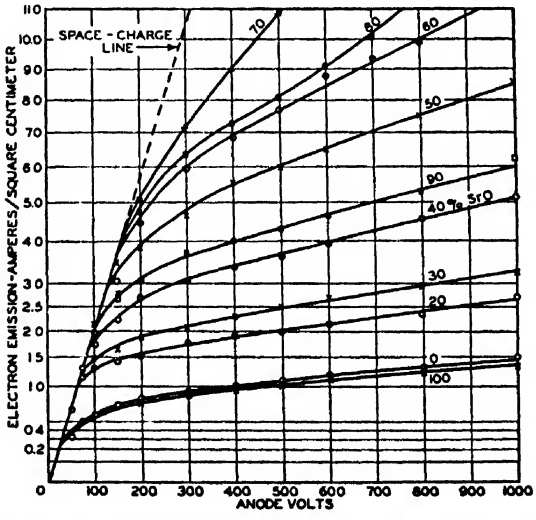


FIG. 4. Electron emission for various coating compositions. Space charge plot. $(I/\mathrm{cm}^2)^{\frac{1}{2}}$ *versus* anode volts.

[8] R. L. Sproull, Phys. Rev. **67**, 166 (1945).
[9] E. A. Coomes, J. App. Phys. **17**, 647 (1946).

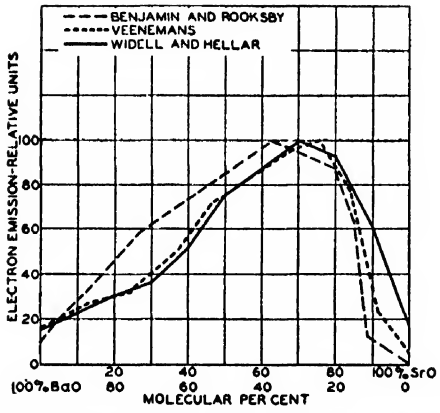


FIG. 5. Comparison of the results of different investigators.

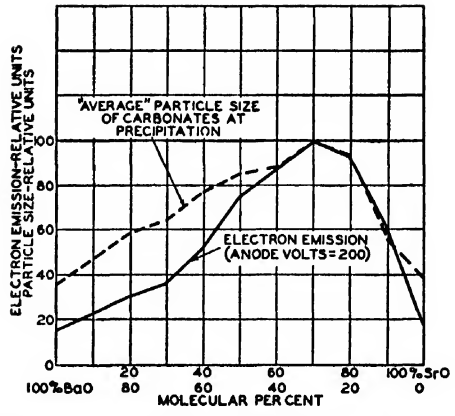


FIG. 6. Comparison of carbonate particle size and electron emission with coating composition.

slumping emission when measured under steady d.c. testing conditions.

DISCUSSION

The Effect of Oxide Ratios on Electron Emission

Figure 2 shows the variation in electron emission for various percentages of BaO and SrO and at various anode voltages. At saturation the curves peak at 70 percent SrO–30 percent BaO (molecular percent). The emission for 100 percent BaO is lower than that observed in previous tests on similar cathodes but this difference may partly be accounted for by the 25° lower operating temperature of this set as shown in Table I. A previous test by Wolfgang Panofsky* of this laboratory showed a much flatter curve with the emission from the BaO about 80 percent of that from the 50–50 ratio but the emission values for the higher percentages of BaO were not stable and decreased during operation.

Figure 4 shows the emission data plotted according to the $\frac{3}{2}$-power law. The large deviations from the normal in the Schottky region for the very highest emission values are due to heating up of the cathodes which could have been lessened if a shorter pulse duration and a lower pulse repetition rate were used.

Comparison with Previous Investigations

When the data obtained in this investigation are compared with those of previous investigators, as shown in Fig. 5, emission values were selected at an anode voltage of 200 volts (about 6500 volts/cm at the cathode) because at that voltage all of the cathodes approached emission saturation. A better comparison might have been made if all investigators had reported zero field emission. Values of emission are given in terms of percent of maximum to facilitate comparison with the published curves of Benjamin-Rooksby and of Veenemans. All three curves have approximately the same general shape except that the Benjamin-Rooksby curve peaks at 63 percent SrO–37 percent BaO while the curve from this investigation peaks at 70 percent $SrCO_3$–30 percent $BaCO_3$.

The Effect of Particle Size on Electron Emission

There has been some discussion in the literature as to the effect of crystal size and particle size on electron emission from oxide-coated cathodes. The change in particle size with variation in ratios of the carbonates when precipitated under identical conditions was measured and the data are shown in comparison with electron emission data in Fig. 6. The curve for the particle size of the carbonates peaks at the same molecular ratio which yields maximum electron emission from the oxide. Although the agreement between the two curves may be merely a coincidence, this subject is being further investigated.

CONCLUSIONS

Maximum electron emission from oxide-coated cathodes consisting of solid solutions of BaO and SrO occurs when the oxides have a molecular percentage of 70 percent SrO and 30 percent BaO.

The maximum size of the precipitated carbonate crystal in the system of $BaCO_3$—$SrCO_3$ occurs at the same molecular ratio which gives the maximum electron emission from the oxides.

There is a close relationship between the size of the carbonate particles and the size of their crystallites as shown by comparing the measurements obtained by the Andreasen pipette with those obtained by the low angle scattering of x-rays on the 70–30 and the 30–70 ratios of carbonates.

The tubes of this test exhibited no measurable electron-emission decay during a 10-microsecond pulse duration.

ACKNOWLEDGMENTS

The authors wish to acknowledge their indebtedness to Miss Irene Kakascik for making the electron micrographs, and to Mr. R. B. DeLong for making the determinations of crystal size by low angle scattering of x-rays.

*Now at California Institute of Technology, Pasadena, California.

Surface Waves and Their Application to Transmission Lines

GEORG GOUBAU
Radio Communication Branch, Coles Signal Laboratory, Fort Monmouth, New Jersey
(Received March 10, 1950)

In this paper the applicability of non-radiating surface waves for transmission lines is investigated. Two types of waves are considered. The first one, originally studied by A. Sommerfeld, is guided by a cylindrical conductor of finite conductivity. Although this wave type has (under comparable conditions) much lower attenuation than the waves in coaxial cables or rigid wave guides, its practical application is restricted by the fact that the extension of the field is very large. Efficient excitation and undisturbed propagation of this wave mode are feasible only for very high frequencies. The other wave type considered in this paper has not been treated in the literature. It is guided by a conductor which is coated with a dielectric layer or the surface of which is otherwise modified; for example, by being threaded. The field of this wave type has a structure similar to that of Sommerfeld's wave, but the extension of the field can be controlled by the surface modification. Thus low loss transmission lines on the basis of this wave type become feasible for frequencies above 100 megacycles. The information necessary for the design of such lines is given and the agreement between the theoretically expected transmission losses and the measured transmission losses is checked.

I. INTRODUCTION

IN 1899, A. Sommerfeld[1] published a paper about wave propagation along an infinitely long cylindrical wire of finite conductivity. The type of wave he investigated is a non-radiating mode which has had little consideration in the modern literature, although the attenuation is theoretically much smaller than that of waves in coaxial cables. There may have been some doubt about the possibility of generating this wave type, as there is in the case of Zenneck's[2] ground wave in wave propagation along the earth. The latter wave is guided by a plane interface separating a non-conducting medium from a conducting one, while in Sommerfeld's case the interface is cylindrical. Both waves are possible solutions of Maxwell's equations satisfying given boundary conditions. However, the solutions are special ones insofar as they refer to *plane* electro-magnetic waves and thus provide that the power sources are infinitely far removed. There has been much discussion about the physical reality of Zenneck's ground wave and there is no convincing answer as yet. However, recent experiments on wave propagation along cylindrical conductors prove that cylindrical surface waves can be generated with high efficiency, and thus they can be employed for guided power transmission. The wave type actually used in these experiments differed from Sommerfeld's wire wave insofar as the surface of the conductor was threaded or coated with a dielectric layer. Such modifications of the conductor surface were made in order to reduce the radial extension of the field and hence the dimensions of the excitation device. The field structure of the modified wave is similar to that of Sommerfeld's wave. There is, however, one interesting difference. Sommerfeld's wave exists only for a conductor of finite conductivity. If the conductivity were increased without limit, the extension of the field would increase in such a manner that the power carried by a wave of finite amplitude would become infinite. The wave propagated by a conductor with a thin coat of dielectric or with an otherwise modified surface however, would exist also in the case of perfect conductivity. The field is only slightly affected by the conductivity, provided the conductivity is above a certain limit.

In order to eliminate misunderstandings, it shall be mentioned that the surface waves considered here are not related to the usual waves on long wire antennas, which are extensively discussed in the literature.[3]

The subject of this paper is a discussion of cylindrical surface waves with regard to their practical application to transmission lines. Two kinds of waves are investigated: The waves on a conductor having a plain surface (Sommerfeld's wave) and the waves on a conductor having a modified surface, in particular on a conductor with a dielectric coat.

II. SOMMERFELD'S WIRE WAVE

An infinitely long cylindrical conductor of circular cross section is embedded in an infinite homogeneous dielectric. The particular solution of this boundary-value problem given by Sommerfeld,[1] describes a radially symmetric transverse magnetic wave travelling along the cylinder. There are also solutions describing waves with cyclic periodicity around the cylinder. These solutions have been investigated by D. Hondros.[4] Because most of the field energy of these unsymmetrical modes is inside the conductor their attenuation is extremely large. Thus, they are of no practical interest and will be excluded from further discussion.

If the coordinate system shown in Fig. 1 is used, the field components of the symmetric wave travelling in

[1] A. Sommerfeld, Ann. Phys. u. Chemie (Neue Folge) 67–1, 233 (1899).
[2] J. Zenneck, Ann. d. Phys. 23, 846 (1907).
Sufficient information about references 1 and 2 is given in J. A. Stratton, *Electromagnetic Theory*, (McGraw-Hill Book Company, Inc., New York, 1941).
[3] S. Schelkunoff, *Electromagnetic Waves*, (D. Van Nostrand Company, Inc., New York, 1943).
[4] D. Hondros, Ann. d. Phys. 30, 905 (1909).

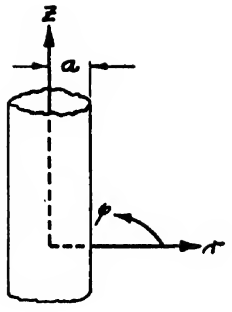


FIG. 1. Coordinate system.

the positive z-direction can be written in the form:

$$E_r = jA\frac{h}{\gamma}Z_1(\gamma r)e^{j(\omega t-hz)},$$
$$E_z = AZ_0(\gamma r)e^{j(\omega t-hz)}, \quad (1)$$
$$H_\varphi = jA\frac{k^2}{\omega\mu\gamma}Z_1(\gamma r)e^{j(\omega t-hz)},$$

as well for the space inside as for the space outside the conductor. The propagation constant h of the guided wave is the same for both parts of the field. The free wave propagation constant k is

$$\text{inside: } k_c=(\omega\mu_c\sigma_c)^{\frac{1}{2}}e^{-j\pi/4}; \quad \text{outside: } k=\omega(\epsilon\mu)^{\frac{1}{2}} \quad (2)$$

(ϵ = dielectric constant, μ = permeability, σ = conductivity). The subscript c refers to the interior of the conductor. The parameters γ and γ_c are defined by

$$\gamma_c^2=k_c^2-h^2; \quad \gamma^2=k^2-h^2. \quad (3)$$

The cylinder functions Z_0 and Z_1 which describe the field inside the conductor are the Bessel functions J_0 and J_1. They are finite at $r=0$. For the outside the Hankel functions $H_0^{(1)}$ and $H_1^{(1)}$ are taken, since they ensure the proper decrease of the field at infinity and give the proper direction of energy flow if the complex parameter γ is in the first quadrant of the complex plane.

The boundary conditions require that E_z and H_φ be continuous at the surface ($r=a$) of the conductor. Thus, we obtain the equations:

$$\mu\frac{\gamma}{k^2}\frac{H_0^{(1)}(\gamma a)}{H_1^{(1)}(\gamma a)}=\mu_c\frac{\gamma_c}{k_c^2}\frac{J_0(\gamma_c a)}{J_1(\gamma_c a)}, \quad (4)$$

$$\gamma^2-\gamma_c^2=k^2-k_c^2=k^2\left(1+j\frac{\sigma\mu_c}{\omega\epsilon\mu}\right)\simeq k^2 j\frac{\sigma\mu_c}{\omega\epsilon\mu}, \quad (5)$$

from which γ and γ_c can be determined and consequently also h. If we restrict our considerations to conductors the radii of which are large compared with the skin depth ($\gamma_c a\gg 1$), J_0 and J_1, in (4) can be replaced by their asymptotic representations:

$$J_0(x)\underset{x\to\infty}{\simeq}\left(\frac{\pi}{2}x\right)^{-\frac{1}{2}}\cos(x-\pi/4);$$
$$J_1(x)\underset{x\to\infty}{\simeq}\left(\frac{\pi}{2}x\right)^{-\frac{1}{2}}\cos(x-3\pi/4). \quad (6)$$

If furthermore the conductor radius is not too large, so that $\gamma a\ll 1$, we can use in (4) the approximations:

$$H_0^{(1)}(x)\underset{x\to 0}{\simeq}\frac{2j}{\pi}\ln(-j0.89x); \quad H_1^{(1)}(x)\underset{x\to 0}{\simeq}-\frac{2j}{\pi x} \quad (7)$$

($0.89\cdots=\frac{1}{2}\times 1.781\cdots=e^C$, where C = Euler's constant $=0.577\cdots$). Equation (4) is then simplified to

$$-\mu\frac{\gamma^2 a}{k^2}\ln(-j0.89\gamma a)=\mu_c\frac{\gamma_c}{k_c^2}\cot(\gamma_c a-\pi/4) \quad (8)$$

where $\cot(\gamma_c a-\pi/4)$ approaches $+j$ since γ_c becomes complex. We shall see later that all these approximations are justified for the considered conductor diameters and frequency range.

Now, (8) can be written in the form:

$$\xi\ln\xi=\eta \quad (9)$$

with

$$\xi=(-j0.89\gamma a)^2,$$
$$\eta=2(0.89)^2\frac{k^2 a\mu_c}{|k_c|\mu}e^{j3\pi/4}=|\eta|e^{j3\pi/4}. \quad (10)$$

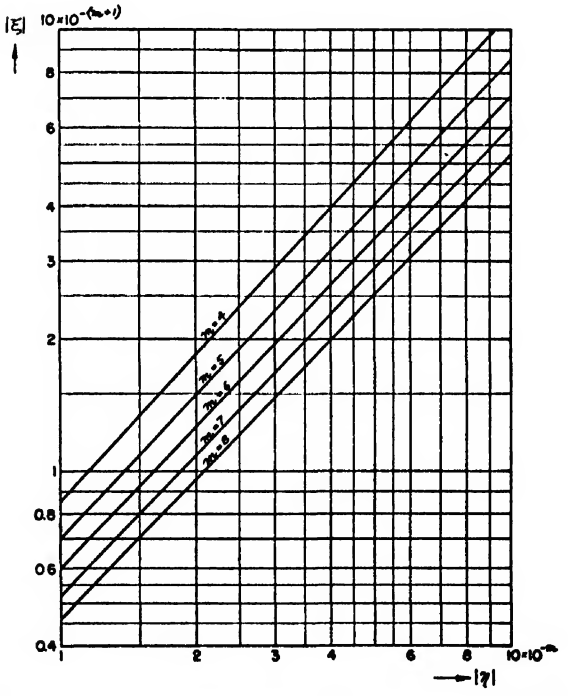


FIG. 2. Graph for determining the absolute value of ξ (see Eq. (12)).

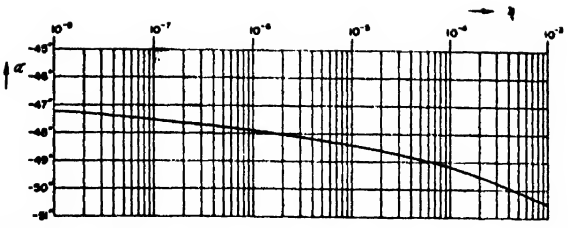


FIG. 3. Graph for determining the argument of ξ (see Eq. (12)).

For copper wire in air

$$|\eta| = 3.27\times10^{-19} af^{\frac{1}{2}} = 1.70\times10^{-3} a\lambda^{-\frac{1}{2}}. \quad (11)$$

f is the frequency and λ the free space wave-length. λ and a are measured in cm. Equation (9) has been used by Sommerfeld for numerical evaluation of γ. With regard to the following discussion about the applicability of these waves for transmission lines, we are concerned with solutions of (9), within a wide range of f and a. If we consider frequencies from 10^8 to 10^{10} cycles/sec. and radii from 0.1 to 1 cm; then, in the case of a copper wire, $|\eta|$ ranges between 3×10^{-8} and 3×10^{-3}. Within this wide range the complex Eq. (9) can be split into two real equations which are more convenient for numerical evaluation. As shown in the Appendix, the absolute value $|\xi|$ and the argument α of $\xi = |\xi| e^{i\alpha}$ can be determined from the equations:

$$|\xi| \ln|\xi| = -|\eta|; \quad \alpha = -\frac{\pi}{4}\left(1 - \frac{1}{\ln|\xi| + 1}\right) \quad (12)$$

$|\xi|$ and α are plotted in Figs. 2 and 3 as functions of $|\eta|$. Each of the 5 curves for $|\xi|$ in Fig. 2 refers to one decade in $|\eta|$, identified by the exponent m. α varies only slightly with $|\eta|$ as seen in Fig. 3.

Having $\xi = |\xi| e^{i\alpha}$, one can determine γ from (10):

$$a\gamma = 1.12(|\xi|)^{\frac{1}{2}} e^{j((\alpha+\pi)/2)}. \quad (13)$$

Within the range of $|\eta|$ considered in Figs. 2 and 3 $|a\gamma| < 1.1\times10^{-2}$, $Re[a\gamma_c] > 15$ and $Im[a\gamma_c] > 5$. Hence the simplifications which lead to Eq. (9) are justified.

The propagation constant h of the surface wave is determined by k and γ (see (3)). As γ/k is very small we can write $h \simeq k - \gamma^2/2k$ or, if γ^2 is expressed in terms of ξ with Eq. (10):

$$h = \left(k + 0.63\frac{|\xi|\cos\alpha}{ka^2}\right) + j\left(0.63\frac{|\xi|\sin\alpha}{ka^2}\right). \quad (14)$$

The real part of h yields the phase velocity of the wave:

$$v = c\left(1 - \frac{0.63|\xi|\cos\alpha}{(ka)^2}\right). \quad (15)$$

The phase velocity of the surface wave is somewhat smaller than the phase velocity c of a free wave in the same dielectric. Consider, for instance, a wave with a frequency of 3000 mc, propagating on a copper wire of 0.1 cm radius, stretched in air. From Eq. (11) we obtain $|\eta| = 5.38\times10^{-6}$ and from Figs. 2 and 3 the corresponding $|\xi| = 3.6\times10^{-7}$ and $\alpha = -48°\ 15'$. Inserting in (16) we find that the reduction in phase velocity is less than 0.004 percent.

The imaginary part of h gives the attenuation β of the wave. β in nepers per cm is

$$\beta = -\frac{0.63|\xi|\sin\alpha}{ka^2}[n/\text{cm}]. \quad (16)$$

The loss L in db per 100 feet becomes:

$$L = -2.66\frac{\lambda}{a^2}|\xi|\sin\alpha\times10^3[db/100\text{ ft.}]. \quad (17)$$

It may be mentioned that β and L are positive because $\sin\alpha$ is negative.

In Fig. 4, L/100 ft. is plotted for various conductor diameters and a frequency range, $10^2 \cdots 10^4$ mc ($\lambda = 3 \cdots 300$ cm). The conductor is assumed to be of copper. It is of interest that even for a wire of only 1 mm radius the losses are considerably smaller than for recommended wave guides. From this fact the application of single conductors for transmission lines appears very promising. However, there is an important question of how to excite efficiently the surface waves. The efficiency of any launching device will be greater the better the agreement between the field built up by the launching device and the field of the wave. The launching device can be considered as a field transformer which transforms the field of a coaxial cable or wave guide into the field of the surface wave. Because the field of the wave extends to infinity, the matching between the field of the launching device and the field of the wave can never be perfect. The larger the launching device the better the chances for good matching conditions. In Section VI we will estimate the loss of a launching device and for this reason we determine in the following that

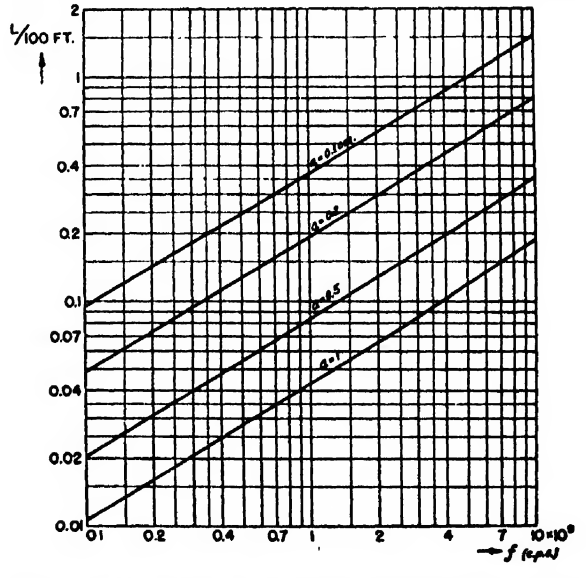


FIG. 4. Transmission loss of copper wires of 0.1, 0.2, 0.5, and 1 cm radius in the case of Sommerfeld's wave.

fraction of the total power of the wave which travels outside of a cylinder of radius ρ surrounding the wire.

The energy flow through the area $\infty > r > \rho$ is given by

$$N_\rho = Re\left[2\pi \int_\rho^\infty r E_r H_\varphi^* dr\right]. \tag{18}$$

(The asterisk denotes the conjugate complex quantity.) The total energy flow N outside the wire is obtained by extending the lower limit of the integral to the wire surface. The quantity we are interested in is N_ρ/N. Inserting for E_r and H_φ the values of Eq. (1) with $Z_0 = H_0^{(1)}$, $Z_1 = H_1^{(1)}$ and considering that $[H^{(1)}(\gamma r)]^* = H^{(2)}(\gamma^* r)$, the power N_ρ at $z=0$ becomes:

$$N_\rho = -Re\left\{2\pi A A^*(\epsilon/\mu)^{\frac{1}{2}} \frac{hk}{\gamma\gamma^*(\gamma^2-\gamma^{*2})} \times[\gamma^*\rho H_1^{(1)}(\gamma\rho) H_0^{(2)}(\gamma^*\rho) - \gamma\rho H_1^{(2)}(\gamma^*\rho) H_0^{(1)}(\gamma\rho)]\right\}. \tag{19}$$

Unfortunately the numerical evaluation of Hankel-functions with complex arguments is cumbersome. However, as we are not too much concerned with high accuracy, we will use the zero representation of the functions up to $|\gamma\rho| < 0.3$, though we have to expect errors of the order of 15 percent for the upper limit of $|\gamma\rho|$. For the wire radii and frequencies considered in Fig. 4, the range $|\gamma\rho| < 0.3$ covers an area around the wire in which the major part of the energy is transmitted.

Using the zero representations of the Hankel functions, shown in (7), and replacing γ by ξ from (10), Eq. (19) becomes:

$$N_\rho = -AA^*(\epsilon/\mu)^{\frac{1}{2}} \frac{ka^4 Re[h]}{|\xi|^2\pi} \times 4(0.89)^4(\ln|\xi|\rho^2/a^2 + \alpha\cot\alpha). \tag{20}$$

The total power N is obtained from (20) by setting $\rho = a$. The expression for N becomes very accurate, since $\gamma a \ll 1$. For the ratio N_ρ/N we get:

$$\frac{N_\rho}{N} \simeq 1 + \frac{2\ln(\rho/a)}{\ln|\xi| + \alpha\cot\alpha}. \tag{21}$$

Within the entire range of $|\xi|$ considered above, α is close to $-\pi/4$. Thus

$$\alpha\cot\alpha \simeq (\pi/4) (\ll |\ln|\xi||)$$

and

$$\frac{N_\rho}{N} \simeq 1 + \frac{2\log(\rho/a)}{\log|\xi| + 0.34} \simeq 1 + \frac{2\log(\rho/a)}{\log 2.2|\xi|}. \tag{22}$$

This equation can be used for determining the radius of that area through which a certain percentage of the power is transmitted. Figure 5 shows the half power radius $\rho_{50\%}$ and the radius for 75 percent power $\rho_{75\%}$ within the same frequency range and for the same wire radii as in Fig. 4. The 75 percent radii are not very accurate for reasons mentioned above. Figure 5 shows that the extension of the field is very large, especially for low frequencies and large wire diameters. For instance, for $f = 1000$ mc/sec., $a = 1$ cm, the half power radius is 28 cm and the 75 percent power radius about 150 cm. A transmission line of this dimension would require a clearance around the conductor of several meters, if frequencies of 1000 mc or less should be transmitted.

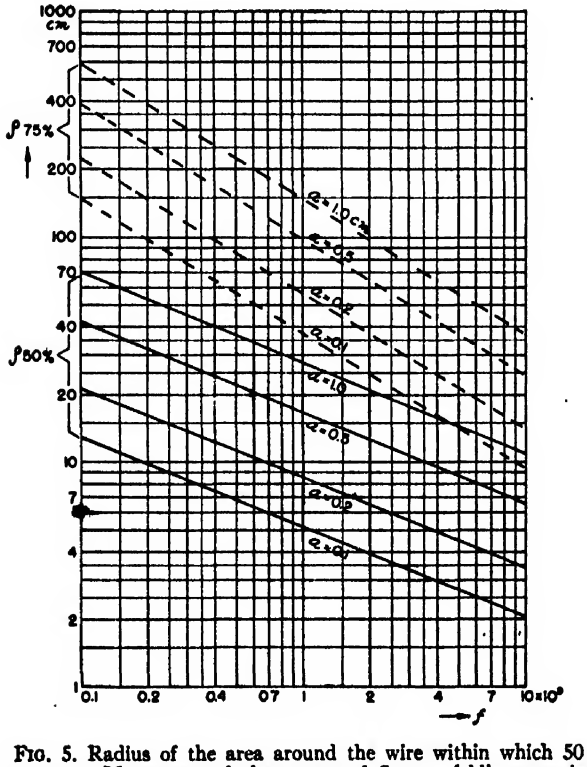


FIG. 5. Radius of the area around the wire within which 50 percent or 75 percent of the power of Sommerfeld's wave is propagated.

In the introduction it was stated that the existence of Sommerfeld's surface wave requires a conductor of finite conductivity. If σ were increased to infinity, $|\eta|$ would become zero and consequently so would $|\xi|$. From (22) we see that N_ρ/N would approach 1 for any finite value of ρ. This means that the power transmitted through a finite area surrounding the wire would be zero if the total power were finite. Thus, the solution of Maxwell's equations loses its sense.

Though a single wire transmission line based on Sommerfeld's surface wave is favorable with regard to the losses effected by the conductor itself, the practical application is restricted to very high frequencies by reason of the large extension of the field. Below 3000 mc the size of low loss launching devices and the clearance area around the wire become too large. It will be shown in the next section how the extension of the field can be

reduced to such an amount that surface wave transmission lines become feasible also for lower frequencies.

III. WAVES ON A CONDUCTOR WITH MODIFIED SURFACE

The waves propagated in coaxial cables or in wave guides differ from Sommerfeld's surface wave in that the finite conductivity of the conductors is an undesired property rather than a necessary condition for the existence of the waves. The question arises: is it possible to modify the surface of a wire so that the conductivity is no longer important for the existence of a surface wave. It is evident that the field of a surface wave on such a guide would not extend as far in the radial direction as the field of Sommerfeld's wave; the latter wave is constrained to the wire only by reason of the finite conductivity while in the other case an additional constraint must be present which concentrates the field around the wire if the conductivity becomes infinite.

In answering the question we first consider the boundary conditions for a surface wave if there is no dissipation present.

An unattenuated surface wave of axial symmetry has field components which are also described by Eq. (1). The cylinder functions in question are the Hankel-functions $H_0^{(1)}$ and $H_1^{(1)}$ with positive imaginary argument. Hence

$$Z_0=H_0^{(1)};\quad Z_1=H_1^{(1)};\quad \gamma=j\gamma';\quad \gamma'^2=h^2-k^2. \tag{23}$$

The Hankel function $H_1^{(1)}$, for positive imaginary argument is negative real and the Hankel function $H_0^{(1)}$, for positive imaginary argument is negative imaginary. Hence E_z has a phase shift of 90° with respect to E_r and H_φ. The boundary conditions at the surface of a guide which propagates a wave of this type demand that E_z and H_φ and therefore their ratio

$$(E_z/H_\varphi)_{r=a'}=\frac{\omega\mu\gamma'}{k^2}\frac{H_0^{(1)}(j\gamma'a')}{H_1^{(1)}(j\gamma'a')}\simeq -j(\mu/\epsilon)^{\frac{1}{2}}\frac{\gamma'^2a'}{k}\ln 0.89\gamma'a' \tag{24}$$

be continuous. a' is the radius at which the internal field and the field of the surface wave meet. Equation (24) states that at the surface of the guide the ratio $(E_z/H_\varphi)_{r=a'}$ must be positive imaginary.

Now we investigate whether the required ratio E_z/H_φ can be obtained by a suitable modification of the surface of a perfectly conducting wire. Consider a cylinder of radius $r=a'$ surrounding a wire which may have any periodic "micro" structure in the z-direction as indicated in Fig. 6. The space between the conductor surface and said cylinder may be partly or wholly filled with dielectric material. This space should be small compared with the wave-length, but large enough compared with the structure on the wire that at $r=a'$ the effect of this structure on the field is averaged. The region bounded

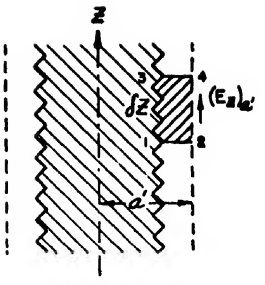


FIG. 6. Conductor with modified surface.

by the cylinder of radius a' is the "guide." With these assumptions the field in the guide can be considered quasi-stationary. The magnetic field is determined by the current I in the wire. The contribution of the electric z-component to the magnetic field can be neglected. The electric field consists in its major part of a potential field originated by the electric charges on the conductor and in its minor part of a curl-field induced by the magnetic field. Current I and charge Q per unit length are connected by the relation

$$\partial I/\partial z+\partial Q/\partial t=0. \tag{25}$$

Since the current distribution propagates in the z-direction with the same phase velocity as the surface wave, (25) can be written in the form

$$-jhI+j\omega Q=0\quad\text{or}\quad Q=hI/\omega. \tag{26}$$

Faraday's induction law applied to the area 1–3–4–2 in an axial plane (see Fig. 6) gives:

$$\delta z\frac{\partial\Phi}{\partial t}=(E_z)_{a'}\delta z+\left(\int_1^2 E_r dr-\int_3^4 E_r dr\right). \tag{27}$$

The length δz, like the distance 1–2, must be small compared with the wave-length. Φ is the magnetic induction flux through a unit length of the area and is proportional to I. In the equation

$$\phi=LI, \tag{28}$$

L can be considered as an inductance per unit length. $(E_z)_{a'}$ in (27) is the longitudinal electric field component at $r=a'$. The integral of the radial electric field component between the points 1 and 2 is proportional to the charge density at 1. A radial induced component of the electric field can be neglected because the dimensions of the roughness of the conductor are assumed to be small compared with the dimensions of the area under consideration. Hence,

$$\int_1^2 E_r dr=Q/C \tag{29}$$

where C has the meaning of a capacitance per unit length. With (29) and (26) the difference between the two integrals in (27) can be written in the form

$$\int_1^2 E_r dr-\int_3^4 E_r dr=-\frac{1}{C}\frac{\partial Q}{\partial z}\delta z=j\frac{h^2}{\omega C}I\delta z. \tag{30}$$

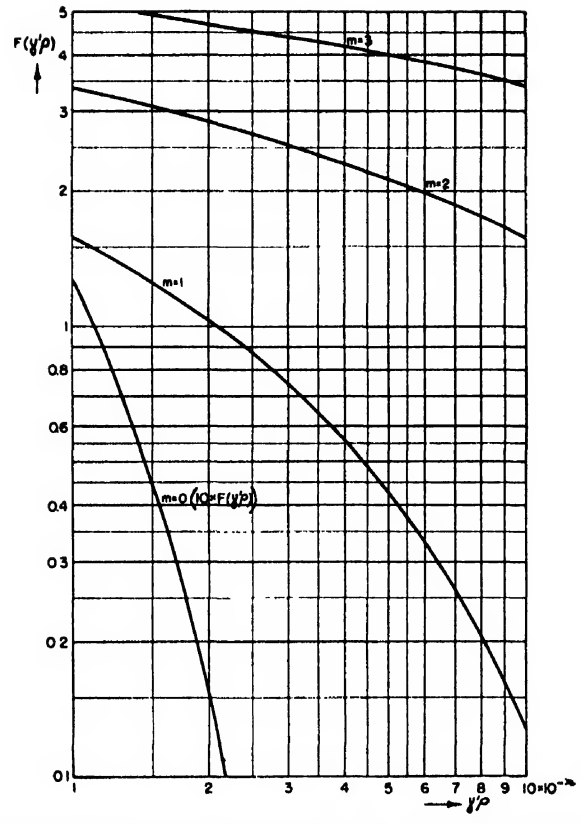


FIG. 7. Graph for determining the extension of the field on surface treated wires (see Eq. (37)).

With (28) and (30) inserted, Eq. (27) becomes:

$$j\omega LI\left(1-\frac{h^2}{\omega^2 LC}\right)=(E_z)_{a'}. \tag{31}$$

Assume for the moment that $(E_z)_{a'}$ is zero. Then, the conditions are like those in a coaxial line having the structure of our guide inclosed in a conductive tube of radius a'. The propagation constant h' of such a line is given by (31) for $(E_z)_{a'}=0$:

$$1-\frac{h'^2}{\omega^2 LC}=0 \quad \text{or} \quad h'^2=\omega^2 LC.^* \tag{32}$$

$\omega^2 LC$ in (31) can now be expressed by h', the propagation constant of a wave which would be translated if the surface wave guide were surrounded by a cylindrical conductor of radius a'. Thus Eq. (31) reads:

$$j\omega LI\left(1-\frac{h^2}{h'^2}\right)=(E_z)_{a'}. \tag{33}$$

As the magnetic field at $r=a'$ is $I/2\pi a'$, the ratio E_z/H_φ at $r=a'$ becomes:

$$(E_z/H_\varphi)_{r=a'}=j2\pi a'\omega L(1-h^2/h'^2). \tag{34}$$

* In truth L and C would change somewhat if the guide were surrounded by a conductor.

The surface wave requires according to (24) an E_z/H_φ having a positive imaginary value. Hence, h' must be greater than k. Because $h=(k^2+\gamma'^2)^{\frac{1}{2}}$ is greater than k, the condition for the existence of the wave considered is that *the guide, inclosed in a conducting cylinder, form a transmission line the phase velocity of which is smaller than the normal velocity in the surrounding dielectric.* Any surface modification of the wire which reduces the phase velocity is suitable for converting the wire into a surface wave guide. Threading of the wire or a dielectric layer on it, for example, accomplishes the required reduction in phase velocity. It may be mentioned that Sommerfeld's wave also is made possible by the fact that the phase velocity is reduced. In that case the reduction is effected by the finite conductivity of the wire. Several steps in the preceding development might require more detailed discussions for their justification. For the sake of brevity such discussions have been omitted.

After having established the existence of a non-radiating surface wave we investigate the extension of the field, still with the assumption that dissipation is negligible. As in the case of Sommerfeld's wave we determine first the amount of power which is transmitted outside a cylinder of radius ρ surrounding the wire. This power is given by (18) and since $H_1^{(1)}(j\gamma' r)$ is real,

$$N_\rho=2\pi AA^*(\epsilon/\mu)^{\frac{1}{2}}\frac{hk}{\gamma'^2}\int_\rho^\infty r[H_1^{(1)}(j\gamma' r)]^2 dr$$

$$=AA^*\pi(\epsilon/\mu)^{\frac{1}{2}}\frac{hk}{\gamma'^2}F(\gamma'\rho) \tag{35}$$

where

$$F(\gamma'\rho)=(\gamma'\rho)^2\left\{-\frac{2}{\gamma'\rho}jH_0^{(1)}(j\gamma'\rho)H_1^{(1)}(j\gamma'\rho)\right.$$

$$\left.-[H_0^{(1)}(j\gamma'\rho)]^2-[H_1^{(1)}(j\gamma'\rho)]^2\right\}. \tag{36a}$$

Hankel functions with imaginary argument are tabulated;[5] thus, the function $F(\gamma'\rho)$ can be calculated easily. In the range $\gamma'\rho<0.1$ the Hankel functions can be approximated by their zero developments with good accuracy. In this way the function $F(\gamma'\rho)$ becomes

$$F(\gamma'\rho)\simeq 8/\pi^2[-\ln 0.89\gamma'\rho-0.5] \quad \text{for} \quad \gamma'\rho<0.1. \tag{36b}$$

The function $F(\gamma'\rho)$ is plotted in Fig. 7. The parameter m indicates the decade in $\gamma'\rho$ to which each curve refers. In the case $m=0$ the function $10\times F(\gamma'\rho)$ is plotted in order to get the curve on the diagram.

The circle of radius ρ_p within which a certain percentage p of the total power of the surface wave travels is given by the equation:

$$p=1-\frac{F(\gamma'\rho_p)}{F(\gamma'a')}. \tag{37}$$

[5] E. Jahnke and F. Emde, *Tables of Functions*, (Dover Publications, New York, 1943).

Figure 8 shows the ratio ρ_p/a as a function of $\gamma' a'$ for $p=50$ percent, 75 percent, 90 percent and 99 percent. For a given diameter a' of the guide the field decreases faster with increasing γ'. Assume, for instance, $a'=1$ cm, $\gamma'=2\times10^{-2}$; then half the power is within a radius of about 6 cm, 75 percent of the power is within $\rho=16$ cm, 90 percent is within $\rho=33$ cm and 99 percent is within $\rho=86$ cm. γ' can be varied by modifying the conductor surface. Thus, we have the possibility of restricting the extension of the field within limits to any desired value. A part of the energy travels in the space between the conductor surface and a', the radius of the guide. But this part is so small within the considered γ' range, that it can be neglected. (See Section IV.)

IV. CONDUCTOR WITH DIELECTRIC COAT

The simplest method for reducing the phase velocity and the extension of the field is the application of a dielectric coat to a plain wire. This case will now be considered in detail. The field in the dielectric layer can be described by Bessel functions (J_0, J_1) and Neumann functions (N_0, N_1). In Eq. (1) Z_0 and Z_1 are to be replaced by

$$Z_0=J_0+bN_0; \quad Z_1=J_1+bN_1, \tag{38}$$

where b is determined from the boundary condition. γ and k are provided with the subscript i indicating that they refer to the insulating layer:

$$\gamma_i^2=k_i^2-h^2; \quad k_i=\omega(\epsilon_i\mu_i)^{\frac{1}{2}}. \tag{39}$$

Let a be the radius of the wire and a' the radius of the outer surface of the dielectric coat; then we obtain from the boundary condition $E_z=0$ at the wire of infinite conductivity:

$$b=-J_0(\gamma_i a)/N_0(\gamma_i a), \tag{40}$$

and the ratio E_z/H_φ at the surface of the coat, $r=a'$, becomes:

$$\left(\frac{E_z}{H_\varphi}\right)_{r=a'}=-j(\mu_i/\epsilon_i)^{\frac{1}{2}}\frac{\gamma_i}{k_i} \times\frac{J_0(\gamma_i a')N_0(\gamma_i a)-J_0(\gamma_i a)N_0(\gamma_i a')}{J_1(\gamma_i a')N_0(\gamma_i a)-J_0(\gamma_i a)N_1(\gamma_i a')}. \tag{41}$$

This equation can be simplified if we are interested only in conditions for which the phase velocity of the surface wave is slightly reduced. These conditions are established if either the dielectric layer is thin compared with the wire radius, or the layer is of the same order of magnitude as the wire radius, but the radius itself is very small compared with the wave-length. In the first case we can write

$$J_0(\gamma_i a)\simeq J_0(\gamma_i a')+\gamma_i J_1(\gamma_i a')(a'-a); \quad N_0(\gamma_i a')\simeq N_0(\gamma_i a')+\gamma_i N_1(\gamma_i a')(a'-a). \tag{42}$$

With these relations (41) becomes:

$$\left(\frac{E_z}{H_\varphi}\right)_{r=a'}\simeq j(\mu_i/\epsilon_i)^{\frac{1}{2}}\frac{\gamma_i^2}{k_i}(a'-a). \tag{43}$$

In the other case we use the zero representations of the cylinder functions. Thus (41) becomes:

$$\left(\frac{E_z}{H_\varphi}\right)_{r=a'}\simeq j(\mu_i/\epsilon_i)^{\frac{1}{2}}\frac{\gamma_i^2}{k_i}a'\ln\frac{a'}{a}. \tag{44}$$

Both Eqs. (43) and (44) are immediately obtained from Eq. (34). If the guide, consisting of the wire and the dielectric layer is thought to be surrounded by a metal tube of radius a', the inductance L per unit length and the propagation constant h' of the coaxial line formed by the guide and the tube are:

$$L=\frac{\mu_i}{2\pi}\ln\frac{a'}{a}; \quad h'=\omega(\epsilon_i\mu_i)^{\frac{1}{2}}=k_i. \tag{45}$$

With these quantities Eq. (34) leads to (44) which we consider more general since it yields, for thin layers, Eq. (43).

At $r=a'$ where the field in the guide and the field of the surface wave meet, $(E_z/H_\varphi)_{r=a'}$ must be continuous. Hence, by comparison of (24) and (44) we get the relation:

$$(\mu/\epsilon)^{\frac{1}{2}}\frac{\gamma'^2}{k}a'\ln 0.89\gamma' a'=-(\mu_i/\epsilon_i)^{\frac{1}{2}}\frac{\gamma_i^2}{k_i}a'\ln\frac{a'}{a} \tag{46}$$

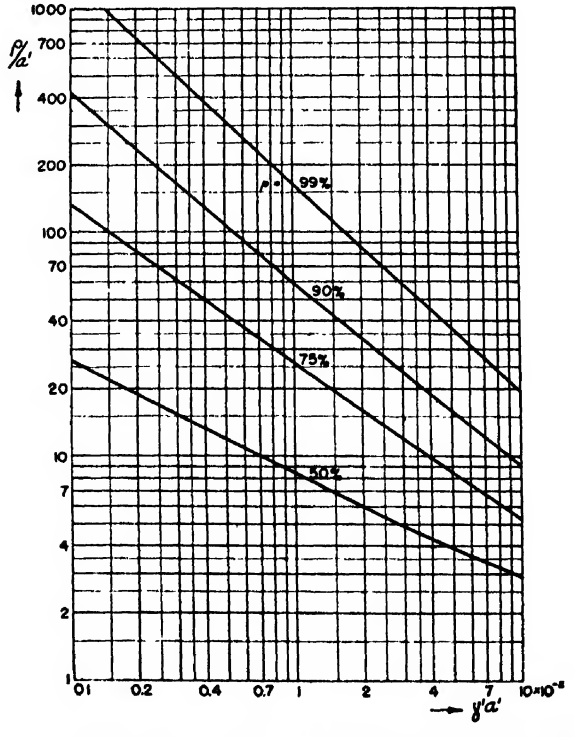


FIG. 8. Relative radius of the area around the wire within which 50 percent, 75 percent, 90 percent and 99 percent of the power is propagated in the case of a conductor with modified surface.

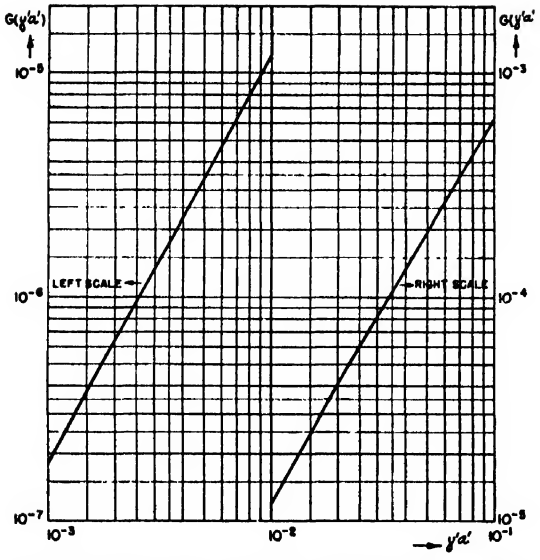


FIG. 9. Design curve in the case of dielectric coated wires (see Eq. (48)).

from which the ratio a'/a is determined:

$$\ln\frac{a'}{a}=-\frac{\epsilon_i}{\epsilon}\frac{\gamma'^2}{\gamma_i^2}\ln 0.89\gamma'a'. \tag{47}$$

As $\gamma_i^2=k_i^2-h^2=k_i^2-k^2-\gamma'^2$ (see Eqs. (39) and (23)) and $\gamma'^2\ll k^2$ Eq. (47) becomes for $\mu_i=\mu$:

$$\ln\frac{a'}{a}\simeq\frac{\epsilon_i}{\epsilon_i-\epsilon}\left(\frac{\lambda}{a'}\right)^2 G(\gamma'a')$$

with

$$G(\gamma'a')=-\left(\frac{\gamma'a'}{2\pi}\right)^2\ln 0.89\gamma'a'. \tag{48}$$

The function $G(\gamma'a')$ is plotted in Fig. 9.

We consider the following example: $a'=1$ cm; $\lambda=50$ cm. The radius within which 90 percent of the power of the surface wave is transmitted, $\rho_{90\%}=50$ cm. What is the required thickness of the dielectric layer if $\epsilon_i/\epsilon=4$?

From Fig. 8 we obtain the required value of $\gamma'a'$ for which $\rho_{90\%}=50$ cm. It is $\gamma'a'=1.21\times10^{-2}$. The corresponding value for $G(\gamma'a')$ is found with Fig. 9 to be 1.68×10^{-5}. Inserted in Eq. (48), these values give the thickness of the layer as 5.6×10^{-2} cm.

The next question to be answered is, how much power is transmitted inside the guide, i.e., within the dielectric coat. This power is determined by E_r and H_φ in the dielectric which are, in terms of the current I,

$$E_r=\frac{Q}{2\pi r\epsilon_i}=\frac{h}{2\pi\epsilon_i\omega r}I \text{ (see (26))}$$
$$H_\varphi=I/2\pi r. \tag{49}$$

Hence:

$$N_i=2\pi\int^{a'} rE_rH_\varphi^*dr=\frac{h}{2\pi\omega\epsilon_i}\ln\frac{a'}{a}II^*. \tag{50}$$

The power N of the surface wave itself is given by (35) with $\rho=a'$. For $F(\gamma'a')$ the approximation (36b) can be used; thus

$$N\simeq-\frac{8}{\pi}AA^*(\epsilon/\mu)^{\frac{1}{2}}\frac{hk}{\gamma'^4}(\ln 0.89\gamma'a'+0.5). \tag{51}$$

The amplitude A can be expressed in terms of the current. Since $(H_\varphi)_{r=a'}=I/2\pi a'$ we find from the field Eq. (1) with the substitutions of (23),

$$AA^*\simeq II^*\frac{\mu\gamma'^4}{16\epsilon k^2}. \tag{52}$$

With (50), (51), (52), (47) and (23) the ratio between N_i and N becomes:

$$\frac{N_i}{N}=-\frac{\epsilon}{\epsilon_i}\frac{\ln(a'/a)}{\ln 0.89\gamma'a'+0.5}$$
$$=\frac{h^2-k^2}{k^2(\epsilon_i/\epsilon)-h^2}\frac{\ln 0.89\gamma'a'}{\ln 0.89\gamma'a'+0.5}. \tag{53}$$

Within the range of $\gamma'a'$ between 10^{-3} and 10^{-1} the term $(\ln 0.89\gamma'a')/(\ln 0.89\gamma'a'+0.5)$ is smaller than 1.25. If we consider reductions in phase velocity of less than 3 percent and assume $\epsilon_i/\epsilon>2.4$, N_i/N is smaller than 6 percent. Thus N_i can be neglected in most practical cases.

V. ATTENUATION ON DIELECTRIC COATED CONDUCTORS

All preceding relations derived for a surface wave guide consisting of a conductor with modified surface are valid under ideal conditions where no dissipation is present. The effect of dissipation can be calculated in the usual manner with the assumption that the field distribution in an equiphase plane is approximately the same as in the dissipationless case.* This means that the losses are calculated for the undisturbed field. The power loss within the path from z to $z+dz$ is proportional to the total power

$$d\bar{N}=-2\beta\bar{N}dz, \tag{54}$$

where β is the attenuation factor in neper/cm.

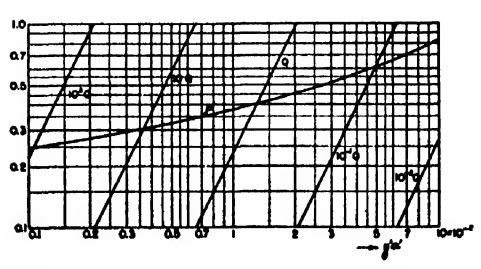


FIG. 10. Chart for determining P and Q (for Eq. (60)).

* This method could not be used in the case of Sommerfeld's wave for obvious reasons.

We consider only the case of a conductor with a dielectric layer. Then the loss consists of two parts, the conductivity loss and the dielectric loss. The conductivity loss, expressed in terms of the current I in the wire, is given by:

$$d\bar{N}_c = -\frac{1}{2\pi a}\left(\frac{\omega\mu_c}{2\sigma_c}\right)^{\frac{1}{2}} II^* dz \tag{55}$$

where μ_c and σ_c are the permeability and the conductivity of the wire. The dielectric loss is

$$d\bar{N}_i = -\frac{h^2}{2\pi\omega\epsilon_i}\ln\frac{a}{a}\tan\delta\, II^* dz$$

$$\simeq -\frac{h^2\gamma'^2}{2\pi\omega\epsilon\gamma_i^2}\ln(0.89\gamma' a')\tan\delta\, II^* dz. \tag{56}$$

$\tan\delta$ is the loss factor of the dielectric material. The total power $\bar{N} = N + N_i$ is obtained from (51)-(53):

$$\bar{N} = -\frac{1}{2\pi}(\mu/\epsilon)^{\frac{1}{2}}\frac{h}{k}\left[\left(1+\frac{\gamma'^2}{\gamma_i^2}\right)\ln 0.89\gamma' a' + 0.5\right] II^*. \tag{57}$$

Equations (54)-(57) give the attenuation factor β_c caused by the conductivity losses and the attenuation factor β_i caused by the dielectric losses. Since $(\gamma'/\gamma_i)^2 \ll 1$, $h \simeq k$

$$\beta_c \simeq -\frac{1}{2a}\left(\frac{\omega\epsilon\mu_c}{2\sigma_c\mu}\right)^{\frac{1}{2}}\frac{1}{\ln 0.89\gamma' a' + 0.5}\ [\text{neper/cm}], \tag{58}$$

$$\beta_i \simeq \frac{1}{2}\frac{\epsilon}{\epsilon_i - \epsilon}\frac{\gamma'^2}{k}\left(1 - \frac{0.5}{\ln 0.89\gamma' a' + 0.5}\right) \times \tan\delta\ [\text{neper/cm}]. \tag{59}$$

With (58) and (59) the total transmission loss L in db per 100 feet becomes

$$L_{100\text{ ft.}} = \frac{P(\gamma' a')}{a(\lambda)^{\frac{1}{2}}} + \frac{\epsilon}{\epsilon_i - \epsilon}\tan\delta\frac{\lambda}{a'^2}Q(\gamma' a')[db] \tag{60}$$

with

$$P(\gamma' a') = -1.33\times 10^4(\epsilon/\mu)^{\frac{1}{2}}\left(\frac{\pi\mu_c}{\sigma_c\mu}\right)^{\frac{1}{2}}\frac{1}{\ln\gamma' a' + 0.38},$$

$$Q(\gamma' a') = 2.11\times 10^3\left(1 - \frac{0.5}{\ln\gamma' a' + 0.38}\right)(\gamma' a')^2. \tag{61}$$

For a coated copper wire stretched in air,

$$P(\gamma' a') = 1.60\frac{1}{-\ln\gamma' a' - 0.38}. \tag{62}$$

The functions $P(\gamma' a')$ (for copper) and $Q(\gamma' a')$ are plotted in Fig. 10.

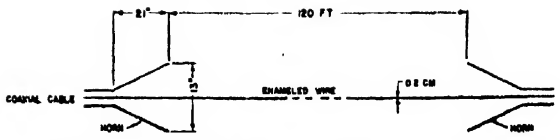


FIG. 11. Sketch of a surface wave transmission line.

For the example considered in the preceeding section ($\lambda = 50$ cm, $\gamma' a' = 1.21\times 10^{-1}$, $a = 0.94$ cm, $a' = 1.00$ cm, $\epsilon_i/\epsilon = 4$) we obtain from Fig. 10: $P = 0.38$, $Q = 0.29$. The loss due to the conductivity is given by the first term in (60) and is 0.057 db/100 ft. Assuming $\tan\delta = 10^{-3}$, the dielectric loss, given by the second term of (60), is 0.005 db/100 ft. Thus the total loss is 0.062 db/100 ft. As may be seen from Fig. 4 for a plain wire of the same radius guiding a Sommerfeld surface wave, the loss would be only half as large. However, the extension of the field would be so high that the conditions for undisturbed propagation and efficient excitation could not be achieved. Figure 5 shows that for $a = 1$ cm and $\lambda = 50$ cm, $\rho_{50\%} = 34$ cm and $\rho_{75\%} = 200$ cm. The corresponding radii for the coated wire given in Fig. 8 ($\gamma' a' = 1.21\times 10^{-1}$) are $\rho_{50\%} = 7.6$ and $\rho_{75\%} = 22.5$ cm; this means they amount to only 22 percent and 11 percent of the other radii.

VI. LAUNCHING OF SURFACE WAVES

In order to excite the desired wave mode with high efficiency a launching device is required which develops within a cross section a field having a structure similar to that of the surface wave. The structure of the fields of both types of surface wave considered in this paper is, within a large area around the conductor, somewhat like in a coaxial cable. The radial electric field component and the circular magnetic field component decrease with $1/r$ as can be seen from the zero representation of the Hankel function $H_1^{(1)}$. Furthermore the ratio E_r/H_φ is very close to $(\mu/\epsilon)^{\frac{1}{2}}$. As the major part of the energy of the waves travels within the area in which the $1/r$ decrease of the field strength is valid, the waves can be launched with good efficiency by means of the device sketched in Fig. 11. The outer conductor of the coaxial feed line is gradually expanded in the form of a horn and the inner conductor of the line is connected to the surface wave guide. The device can be considered as a tapered coaxial line. The arrangement on the receiving end is the same. In the case of Sommerfeld's surface wave the area in which the field should be simulated is very large especially for lower frequencies. Hence, the efficient excitation of this wave would require an extremely large horn. In the other case, where the extension of the field is reduced by means of a suitable modification of the conductor surface, the dimensions of launching devices can be kept within reasonable limits.

The efficiency of the launching device of Fig. 11 can be determined approximately if the opening of the horn does not too much exceed the area within which the field decreases with $1/r$. In this case, considering the receiving

end, that portion of the wave will be received which falls within the area of the horn. If the radius of the horn is equal to $\rho_{75\%}$, for instance, about 75 percent of the energy will be received. Because of the reciprocity theorem the efficiency of the horns for transmission and reception must be the same.

The tapering of the horns is important too, insofar as in the case of large angles higher modes may develop inside the horns, and besides, the wave front at the end of the horns may deviate too much from the required plane wave front. It is desirable to have the surface modification started within the horns, in order to have the surface wave build up therein.

In the case of a surface wave guide consisting of a wire with modified surface, the efficiency of a horn can be determined with the aid of Fig. 7. If the parameter γ' and the radius a' of the guide are known $F(\gamma'a')$ can be obtained. Let b be the radius of the horn at the open end; then the efficiency is given by $1-F(\gamma'b)/F(\gamma'a')$ where $F(\gamma'b)$ is also obtained from Fig. 7.

VII. EXAMPLES

The set-up considered here is shown in Fig. 11. The surface wave guide consists of an enameled wire of 120-foot length having a diameter of 0.2 cm ($a=0.1$) and an enamel coat of 0.005 cm thickness. The dielectric constant and the power factor of the enamel are $\epsilon_i/\epsilon=3$ and $\tan\delta=8\times10^{-3}$, for low frequencies. $\tan\delta$ will probably be much larger for the high frequencies under consideration; however, having no better data available, we will use the quoted figure. The diameter of the horns is $2b=33$ cm.

First we determine the loss of the wire, assuming a frequency of 3300 mc. With (48) we find, $G(\gamma'a')=4.04\times10^{-6}$. From Fig. 9 we obtain the corresponding $\gamma'a'=5.6\times10^{-3}$; Fig. 10 gives for $P(\gamma'a')$ and $Q(\gamma'a')$ the values $P=0.33$, $Q=0.072$. These values inserted in (60) deliver the transmission loss per 100 feet as 1.35 db. Hence the loss for 120 feet is 1.62 db.

In order to determine the loss due to the launching device we determine $F(\gamma'a')$ and $F(\gamma'b)$ from Fig. 7, $F(\gamma'a')=4.9$, $F(\gamma'b)=0.17$. The efficiency of each horn is then 95.6 or the loss is 0.20 db. As both horns together contribute 0.40 db, the total theoretical loss of the transmission line becomes 2.0 db. The measured loss was 2.3 db. It should be mentioned that the loss of the setup does not vary more than 0.1 db within the frequency range from 1500 to 3400 mc.

As another example we consider a transmission line with the same launching device and an enameled wire of 600-foot length and 0.322 cm. diameter. The thickness of the insulation is 0.0025 cm. We calculate the loss for a frequency of 1600 mc. From (48) we obtain $G(\gamma'a')=0.75\times10^{-7}$ and from Fig. 9, $\gamma'a'=2.4\times10^{-3}$. The corresponding values for $P(\gamma'a')$ and $Q(\gamma'a')$ are found in Fig. 10. They are $P(\gamma'a')=0.28_3$ and $Q(\gamma'a')=0.013_1$. If these values are inserted in (60) the loss for the length of 600 feet is 2.65 db. The loss contributed by one horn is obtained with Fig. 7 which gives the values $F(\gamma'a')=4.6$ and $F(\gamma'b)=0.88$. Hence the efficiency is 80.9 percent or the loss 0.9_3 db. As the two horns contribute 1.8_6 db, the total loss of the transmission line becomes 4.5 db. The measured loss was approximately 5 db. It may be mentioned that the wire was supported at 100 feet intervals by strings about 6 feet above the ground.

VIII. ACKNOWLEDGMENT

Thanks are due Mr. A. Meyerhoff for his invaluable assistance in the preparation of the manuscript and Mr. C. Sharp and Mr. L. Battersby for valuable suggestions and for their cooperation in the performance of the experiments of which part of the results are quoted in this paper.

IX. APPENDIX

Derivation of Eq. (12):

If we write for ξ and η

$$\xi=|\xi|e^{j\alpha}=|\xi|e^{-j(\pi/4+x)};\quad \eta=|\eta|e^{j3\pi/4}\ (-\pi<x<\pi) \tag{A1}$$

(9) becomes

$$|\xi|e^{-jx}\left[\ln|\xi|-j\left(\frac{\pi}{4}+x\right)\right]=-|\eta|. \tag{A2}$$

This equation can be split into two equations if the real and imaginary parts are separated:

$$|\xi|\ln|\xi|\cos x\cdot\left(1-\frac{\pi/4+x}{\ln|\xi|}\tan x\right)=-|\eta|; \tag{A3}$$

$$\tan x=-\frac{\pi/4+x}{\ln|\xi|}. \tag{A4}$$

Inserting (A4) into (A3),

$$|\xi|\ln|\xi|\cos x\cdot(1+\tan^2 x)=-|\eta| \quad\text{or}\quad \frac{|\xi|\ln|\xi|}{\cos x}=-|\eta|. \tag{A5}$$

Within the range considered, $|\eta|<10^{-3}$, and $|\xi|\ll1$ (because $\gamma\ll k$). As $|\cos x|$ is certainly ≤1, $|\xi|$ will be smaller than 1.1×10^{-4} and $-\ln|\xi|>9$. From (A4) we see that for $-\ln|\xi|>9$, x becomes small:

$$x\simeq\frac{\pi}{4}\frac{1}{-\ln|\xi|-1}\left(<\frac{\pi}{32}\right). \tag{A6}$$

For such small angles $\cos x$ in (A5) can be replaced by 1. Thus the Eqs. (12) are obtained.

A Note on Control Area

THOMAS M. STOUT
Electrical Engineering Department, University of Washington, Seattle, Washington
(Received May 29, 1950)

Control area, or the integrated error for a step function input, has been used as a measure of control system performance. For systems described by

$$\frac{\theta_0(s)}{\theta_i(s)}=e^{-sT_d}\frac{(\tau_1 s+1)(\tau_2 s+1)\cdots(\tau_m s+1)}{(T_1 s+1)(T_2 s+1)\cdots(T_n s+1)},$$

it is proved that the control area is

$$A=T_d+\sum_{i=1}^{n}T_i-\sum_{i=1}^{m}\tau_i.$$

A simple method of evaluation is presented and some limitations of the control area criterion are mentioned.

A COMMON measure of the response of a control system[1–3] to a unit step function is the "control area," shown as the shaded area in Fig. 1. The control area is defined mathematically as the integral of the error with respect to time from zero to infinity, the error being defined as the difference between input and output. An optimum adjustment of the system parameters is obtained when the control area is a minimum.

Evaluation of the control area in terms of the system parameters by conventional methods is a difficult task if the system is complicated. However, for systems described by the equation

$$\frac{\theta_0(s)}{\theta_i(s)}=\frac{(\tau_1 s+1)(\tau_2 s+1)\cdots(\tau_m s+1)}{(T_1 s+1)(T_2 s+1)\cdots(T_n s+1)} \tag{1}$$

$$=\frac{P(s)}{Q(s)}, \tag{2}$$

subject to a step function of magnitude M, it will be shown that the control area is simply

$$A=1/M\int_0^\infty e(t)dt$$

$$=(T_1+T_2+\cdots+T_n)-(\tau_1+\tau_2+\cdots+\tau_m). \tag{3}$$

In these equations, θ_0, θ_i, and e are standard symbols[4] for the output, input, and error, respectively, functions of s are the Laplace transforms of the corresponding functions of time, and $\tau_1\cdots\tau_m$, $T_1\cdots T_n$ are time constants of the system. Equation (1) describes systems with any number of "lag" or "lead" (delay or anticipatory) elements having zero steady-state error for a step function input. If the steady-state error is not zero, the control area as defined above is infinite and therefore useless as a measure of system performance.

[1] R. C. Oldenbourg and H. Sartorius, *The Dynamics of Automatic Controls* (A.S.M.E., New York, 1948), pp. 65–67.
[2] C. Mack, Phil. Mag. **40**, 922–928 (September, 1949).
[3] D. P. Eckman, *Principles of Industrial Process Control* (John Wiley and Sons, Inc., New York, 1945), p. 161.
[4] G. S. Brown and D. P. Campbell, *Principles of Servomechanisms* (John Wiley and Sons, New York, 1948).

PROOF

By means of Eq. (2), the error may be expressed in terms of the input by the equation

$$e(s)=\left[\frac{Q(s)-P(s)}{Q(s)}\right]\theta_i(s). \tag{4}$$

If the input is a unit step function whose Laplace transform is $1/s$, the transform of the error is

$$e(s)=\frac{Q(s)-P(s)}{sQ(s)}. \tag{5}$$

Denote by $I(t)$ the indefinite integral of the error

$$I(t)=\int e(t)dt. \tag{6}$$

Since $I(0)$ is necessarily zero, the Laplace transform of $I(t)$ is

$$I(s)=e(s)/s. \tag{7}$$

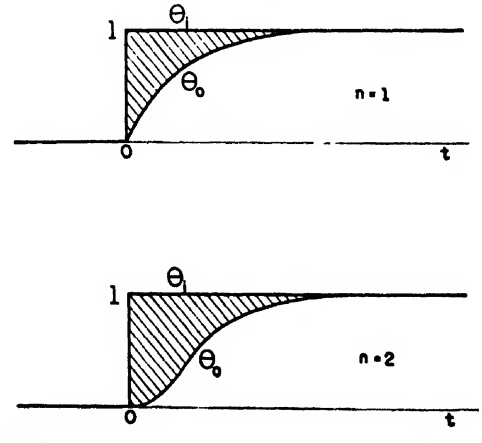


FIG. 1. Response and control area for simple control systems.

The control area may now be found from Eq. (7) by means of the final value theorem,[5,6] giving

$$A=\lim_{t\to\infty}[I(t)]$$
$$=\lim_{s\to 0}[sI(s)]$$
$$=\lim_{s\to 0}[e(s)]. \tag{8}$$

Substituting the expression for $e(s)$ from Eq. (5) into Eq. (8) gives

$$A=\lim_{s\to 0}\left[\frac{Q(s)-P(s)}{sQ(s)}\right], \tag{9}$$

which is indeterminate, since $Q(0)=P(0)=1$. Differentiating numerator and denominator, we obtain

$$A=\lim_{s\to 0}\left[\frac{Q'(s)-P'(s)}{sQ'(s)+Q(s)}\right] \tag{10}$$
$$=\lim_{s\to 0}[Q'(s)-P'(s)]. \tag{11}$$

At this point it will be convenient to write

$$P(s)=\prod_{i=1}^{m}p_i$$
$$Q(s)=\prod_{i=1}^{n}q_i, \tag{12}$$

where $p_i=\tau_i s+1$ and $q_i=T_i s+1$. The derivative of $P(s)$ or $Q(s)$ with respect to s is the sum of all the products formed by multiplying the derivative of one factor by all the other factors, or

$$P'(s)=\sum_{j=1}^{m}\tau_j\Big[\prod_{i=1}^{j-1}p_i\Big]\Big[\prod_{j+1}^{m}p_i\Big]$$
$$Q'(s)=\sum_{j=1}^{n}T_j\Big[\prod_{i=1}^{j-1}q_i\Big]\Big[\prod_{j=1}^{m}q_i\Big]. \tag{13}$$

Since p_i and q_i approach one as s approaches zero, we get

$$\lim_{s\to 0}P'(s)=\tau_1+\tau_2+\cdots+\tau_m$$
$$\lim_{s\to 0}Q'(s)=T_1+T_2+\cdots+T_n \tag{14}$$

and the proof of Eq. (3) is complete.

In systems characterized by "dead time," the output response does not begin immediately but starts at a later time, $t=T_d$. In this case,

$$\frac{\theta_0(s)}{\theta_i(s)}=\frac{e^{-sT_d}P(s)}{Q(s)} \tag{15}$$

and a similar proof will show that now

$$A=T_d+\sum_{i=1}^{n}T_i-\sum_{i=1}^{m}\tau_i. \tag{16}$$

This conclusion is also evident from the previous result upon consideration of the nature of the response.

DISCUSSION

Using conventional methods, the error would be written

$$e(t)=\sum_{i=1}^{m}K_i e^{-t/T_i}, \tag{17}$$

where the coefficients K_i are functions of $\tau_1\cdots\tau_m$ and $T_1\cdots T_n$, and the control area found by direct integration. The coefficients may be determined from the initial conditions* or from Eq. (5) by the use of partial fractions or residues. For the simple first- and second-order systems, whose responses are shown in Fig. 1, the error is

$$e_1(t)=e^{-t/T_1} \tag{18}$$

$$e_2(t)=\frac{T_1}{T_1-T_2}e^{-t/T_1}+\frac{T_2}{T_2-T_1}e^{-t/T_2}. \tag{19}$$

In these simple cases, the validity of Eq. (3) is easily demonstrated. However, the algebraic difficulties encountered in treating the general case by this method are staggering; the direct attack made possible by the Laplace transformation is simpler and more successful.

If $P(s)$ and $Q(s)$ should be given in the form

$$P(s)=B_m s^m+B_{m-1}s^{m-1}+\cdots+B_1 s+1$$
$$Q(s)=A_n s^n+A_{n-1}s^{n-1}+\cdots+A_1 s+1, \tag{20}$$

the control area can be found directly without solving for the individual time constants. Expansion of the factored forms of $P(s)$ and $Q(s)$ used in Eq. (1) will show that

$$B_1=\tau_1+\tau_2+\cdots+\tau_m$$
$$A_1=T_1+T_2+\cdots+T_n. \tag{21}$$

The control area is therefore given by A_1-B_1 directly. Using the control area criterion, the objective of system design would be to minimize (A_1-B_1). It should be remembered that allowable values of A_1 will be fixed by stability considerations, as given by Routh's stability criterion.[6,7] In particular, if $A_1\leqq 0$, the system will be unstable.

[5] R. V. Churchill, *Modern Operational Mathematics in Engineering* (McGraw-Hill Book Company, Inc., New York, 1944).

[6] M. F. Gardner and J. L. Barnes, *Transients in Linear Systems* (John Wiley and Sons, Inc., New York, 1942).

* For a unit step function, the error is equal to one and the first $n-1$ derivatives with respect to t are equal to zero.

[7] E. J. Routh, *Advanced Rigid Dynamics* (Macmillan Company, Ltd., London, 1930).

If the time constants are complex, Eq. (3) is still applicable. In physical systems complex time constants occur in conjugate pairs, such as

$$T=1/(a+j\omega_n),$$
$$\bar{T}=1/(a-j\omega_n),$$

where a is a damping factor and ω_n is the natural frequency. For such a pair of complex time constants,

$$T+\bar{T}=2a/(a^2+\omega_n^2). \quad (22)$$

This result indicates nicely one limitation of control area as a criterion for system design, since the minimum value of $T+\bar{T}$ is obtained when the damping is zero. Other systems, with equally undesirable responses, can be imagined for which the control area might be zero or even negative. These observations indicate the need for simultaneous consideration of the relative stability of the system. In systems for which the control area is a suitable measure of response, the methods of evaluation proposed in this note may be useful.

A Small Electron Microscope

JOHN H. REISNER AND EDMUND G. DORNFELD
Radio Corporation of America, Camden, New Jersey
(Received May 25, 1950)

A new electron microscope of greatly simplified design has been constructed. The instrument has a resolution of 100 Angstroms, yet is small enough to be mounted on a desk or laboratory table. Magnetic lenses are energized by permanent magnets. An accelerating voltage of 50 kilovolts is varied over a narrow range to provide means of focus. Images may be photographed on 2×2 plates. Plates may be changed in a minute, and specimens in fifteen seconds. Direct magnifications of 1500, 3000 and 6000 are possible depending on pole pieces used. Photographic images may be enlarged up to ten times. The instrument is free of x-radiation and completely interlocked and provided with shorting devices for protection from high voltage.

INTRODUCTION

ALL electron microscopes have been characterized heretofore by large size and weight, and usually by complexity of design. These characteristics are a result of extremely great operational flexibility of the instruments. In the early days of electron microscopy, such flexibility was essential to the determination of the best means of utilizing the optical components and to the development of new techniques. Complexity is still a necessary evil of the highest resolution instruments, and of those engaged in an extremely wide range of problems. However, many of the uses to which the electron microscope is now put[1] require only the moderately high resolution which may be realized on comparatively simple instruments.

A first step in the direction of smaller, simpler microscopes was made several years ago,[2] culminating in a well-known commercial instrument.[3] Additional simplification is possible by utilizing permanent magnets to replace coil windings, cables, control circuits, and power supplies. Unfortunately, the use of permanent magnets fixes the magnification for a given set of lens pole pieces. However, at the same time, alignment problems largely disappear so that transverse adjustments may be omitted, along with the attendant vacuum difficulties. By being content with ultimate resolutions of 100 Angstroms or less, the condenser lens may be omitted and its beam control function delegated to an aperture. By limiting direct magnification to 3000 or 6000 times, the column may be shortened, the volume of the vacuum system greatly reduced, and in turn, the vacuum diffusion pump may be reduced in size. Utilization of new materials such as silicone diffusion pump oil, combined with a new treatment of old valving and stage problems, have resulted in the greatly simplified electron microscope shown in Figs. 1 and 2.

DIMENSIONAL CONSIDERATIONS

The first step toward simplification is the reduction of size. The minimum dimensions for the microscope optical system are determined primarily by the highest magnification required, and by the focal lengths of the lenses. For the highest magnification, the lenses are so powerful that the magnification may be expressed to good approximation by the expression $M\simeq p/f$, where "M" is the magnification, and "p" and "f" are the image and focal distances of a lens. It is simply demonstrated that the most favorable utilization of the column length "L" is that where both lenses have the same image distances, i.e., where $p_1=p_2\cong L/2$, since f is very small compared to p. The minimum required column length then becomes $L_{\min}\cong 2(M_t f_1 f_2)^{\frac{1}{2}}$, where "$M_t$" is the total magnification, and "f_1" and "f_2" are the focal lengths of the objective and projector lenses respectively. For the microscope described, $f_1=0.25$ cm, and $f_2=0.20$ cm, when the accelerating voltage is 50 kilovolts and the over-all magnification is 6000 times. Under these circumstances, the minimum length turns out to be about 35 centimeters or 14 inches. It is inter-

[1] R. W. G. Wyckoff, *Electron Microscopy Technique and Applications* (Interscience Publishers, Inc., New York, 1949), pp. 22–25, Chapters VI and VII.
[2] V. K. Zworykin and J. Hillier, J. App. Phys. 14, 658 (1943).
[3] P. C. Smith and R. G. Picard, Radio News 32, 41 (1944).

esting that a saving of 10 centimeters in length could be accomplished by limiting the top direct magnification to 3000 times.

The distance from the specimen to the illuminating source is also limited by magnification. The higher the magnification, the smaller must become the distance in order to maintain illumination. Unfortunately, this distance also influences resolution adversely as it is decreased. The actual values of the distances between optical parts are shown in Fig. 3. It is seen that the distance between the objective and projector lens gaps is substantially smaller than $L/2$. The distance is actually $0.37L$, decreasing the maximum magnification possible in length L by six percent, which is negligible. Design considerations governing the lens energizing system and the mechanical construction of the pole pieces determine the spacing between the two lens gaps.

The short space below the column provided for the opening of the plate holder door, and the space above the column necessary to provide protection against high voltage flash-over, raise the total length of the microscope assembly to 30 inches from base to top of the hood. The diameter of the body of the microscope is determined by the lens. Smaller diameters than 4.75 inches result in the operation of the permanent magnets at too high a demagnetization coefficient.

THE OPTICAL SYSTEM

The optical elements of the microscope are all fixed, with the exception of the electron gun assembly which is externally adjustable. This need for adjustment in the gun system is occasioned by the demountable nature of the filament which must be replaced from time to time. Since filaments differ somewhat in shape and in the position of the point of maximum electron emission, adjustment must be provided for the gun assembly. The gun assembly comprises a filament-grid assembly, rigidly attached to an anode. The anode is positioned with respect to the other optical elements by means of two screws loaded by a return spring, Figs. 4 and 5. The filament-grid portion of the gun is physically described[4] and its operation is discussed[5] in the literature. The electrodes are connected in a self-biased circuit, giving rise to a very intense solid cone of illumination. It is the function of the transverse gun adjustment controls to select that element of the cone of illumination which is parallel to the optical axis of the system and position it on that axis.

The alignment reference of the entire optical system is the bore of the lens energizing spool. The projector pole piece assembly and the lower element of the objective pole piece are positioned in the spool hole to the tolerance of the machining, nominally 0.0003 inch. The position of the upper element of the objective lens must be experimentally determined and fixed when the instrument is initially put in operation. Early embodiments of the instrument positioned the upper lens ele-

FIG. 1. The small electron microscope column completely assembled.

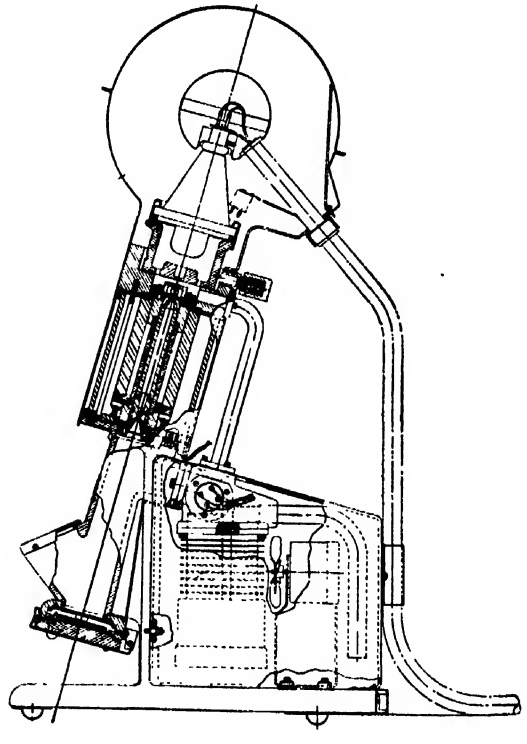


FIG. 2. A cross-sectional view of the electron microscope column employing permanent magnets.

[4] J. H. Reisner and R. G. Picard, Rev. Sci. Inst. 19, 556 (1948).
[5] J. Hillier and S. G. Ellis, J. App. Phys. 20, 700 (1949).

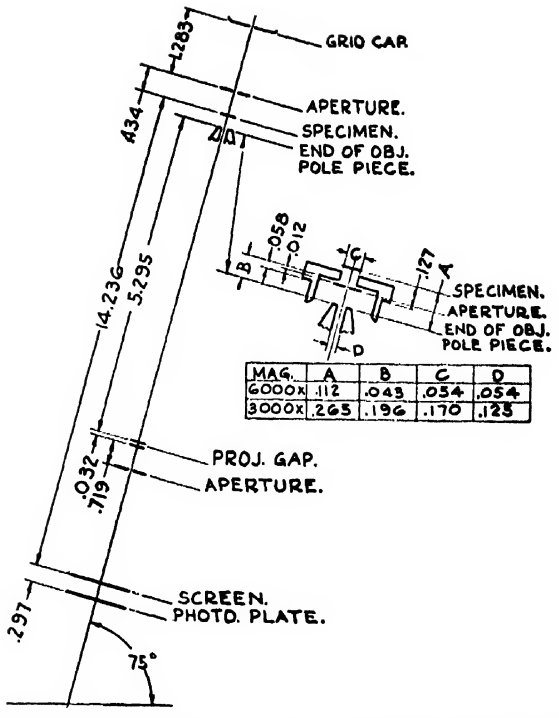
ment by its accurate fit into the energizing spool bore. However, the lateral positioning of the objective pole piece is extremely critical. The nominal build-up of fabrication tolerances amounting to 0.0003 inch can provide an image shift at the final screen of 0.15 inch per kilovolt change in accelerating voltage at fifty kilovolts and at a magnification of 6000 times. So large a shift is objectionable. To permit accurate electrical alignment of the upper objective element, a reference yoke is provided to assure position of that element. The yoke is transversely adjustable by means of four tapered screws (Fig. 5), which also serve to lock it in position so that the lens element may be removed and replaced without affecting alignment. Once aligned, the adjustment for the yoke is correct for all sizes and combinations of objective elements.

The objective aperture is referenced to the hole of the upper objective lens element, and is positioned 0.058 inch below the upper surface of the pole piece. The specimen is positioned 0.023 inch above the upper surface of the pole piece.

An aperture placed 0.434 inch above the specimen is used to exercise the beam control function normally accomplished by the condenser lens system of larger microscopes. Actually, the lack of a condenser lens ultimately limits the resolution of the instrument to that commensurate with the angular aperture of the electron gun. The disk-shaped aperture fits accurately in a hole of a heavy aperture holder, which is in turn referenced to the same yoke as that which aligns the upper lens element of the objective (Fig. 5). The massive aperture holder firmly held in the yoke does not move under the heating from the gun bombardment and shields the specimen rod from the heating and contamination of the gun.

Since permanent magnets energize the magnetic lenses, change in focal length may be accomplished only by changing the accelerating voltage applied to the microscope gun. The range of voltage necessary to provide focus for a wide range of specimens and holders is of the order of 500 volts at 50 kilovolts. A wider range is desirable for checking alignment and for checking the gap spacing of the objective pole pieces. A four-kilovolt swing produced through continuous coarse and a fine control has been found most suitable.

In order to assure the occurrence of a focused image at given voltage, e.g., 49 kilovolts, some variation in one of the focal lengths or object distances must be provided by mechanical means to compensate for differences in the build-up of mechanical tolerances of components and variations in the strength of energizing magnets. This required adjustment is provided in the gap spacing of the objective lens. The upper pole face of the objective pole piece seats against the energizing spool, and is, therefore, fixed with respect to the specimen. The lower pole of the objective, and the complete projector pole piece assembly are rigidly attached to one another. Motion of the lens group along the axis



MAG.	A	B	C	D
6000X	.112	.043	.054	.054
3000X	.265	.196	.170	.125

FIG. 3. A dimensional diagram of the optical system showing lens dimensions for magnifications of 3000× and 6000×.

of the lens energizing spool is possible, and permits adjustment of the spacing of the objective pole piece gap. The axial movement is possible in fine increments with reproducibility by means of the screw thread which advances the projector pole piece as it is turned.

Changing the gap separation shifts the principal planes of the lens in the direction of the change. Since focal length is relatively independent of gap separation, the change in focal length due to gap increase is small compared to shift of the principal plane.[6] The result is that the lens system may be positioned so as to bring the focal plane of the objective into coincidence with the specimen plane for any of the available range of voltages by simply changing the objective gap.

In the initial adjustment of the instrument, too small a gap will result in a focal plane on the gun side of the specimen which manifests itself in an approach toward focus at the low voltage end of the focus controls. Such evidence indicates that the pole piece gap must be increased by turning the projector pole piece counterclockwise. Likewise, approaching focus at the high voltage end of the focus control means that the objective gap is excessive, which must be corrected by a clockwise rotation of the projector pole piece. The amount of the rotation depends upon the objective pole piece in use and upon the accelerating voltage. The required pole piece rotation may be estimated

[6] Zworykin, Morton, Ramberg, Hillier and Vance, *Electron Optics and The Electron Microscope* (John Wiley and Sons, Inc., New York, 1945), p. 153.

from the relation:

$$\frac{1}{f}=\frac{K}{V}$$

taken from the magnetic lens formula, where "f" is the focal length, "V" the accelerating voltage, and "K" a constant depending upon factors fixed for this application. From this relation, it is apparent that:

$$\frac{df}{f}=\frac{dV}{V}$$

In powerful lenses, used at large magnification as herein, it may be assumed that $f\cong q$, the object distance, so that for small increments:

$$\frac{\Delta q}{\Delta V}=\frac{q}{V}.$$

For the 3000× objective of Fig. 6, $q=0.138$ inch, and $V=50{,}000$ volts.

$$\frac{\Delta q}{\Delta V}\cong 0.0028 \text{ mils/volt.}$$

Since the gap change results from turning a 32 thread,

$$\frac{\Delta\theta}{\Delta V}\cong 0\ 033 \text{ degree/volt,}$$

where "θ" is the rotation of the pole piece required to advance it. Hence, if the voltage at which focus occurs is to be increased from 46 to 49 kilovolts, the pole piece must be turned 99 degrees counterclockwise.

The necessity for cleaning off contamination deposited on pole pieces during operation is the reason for the demountable construction of lens elements. Such con-

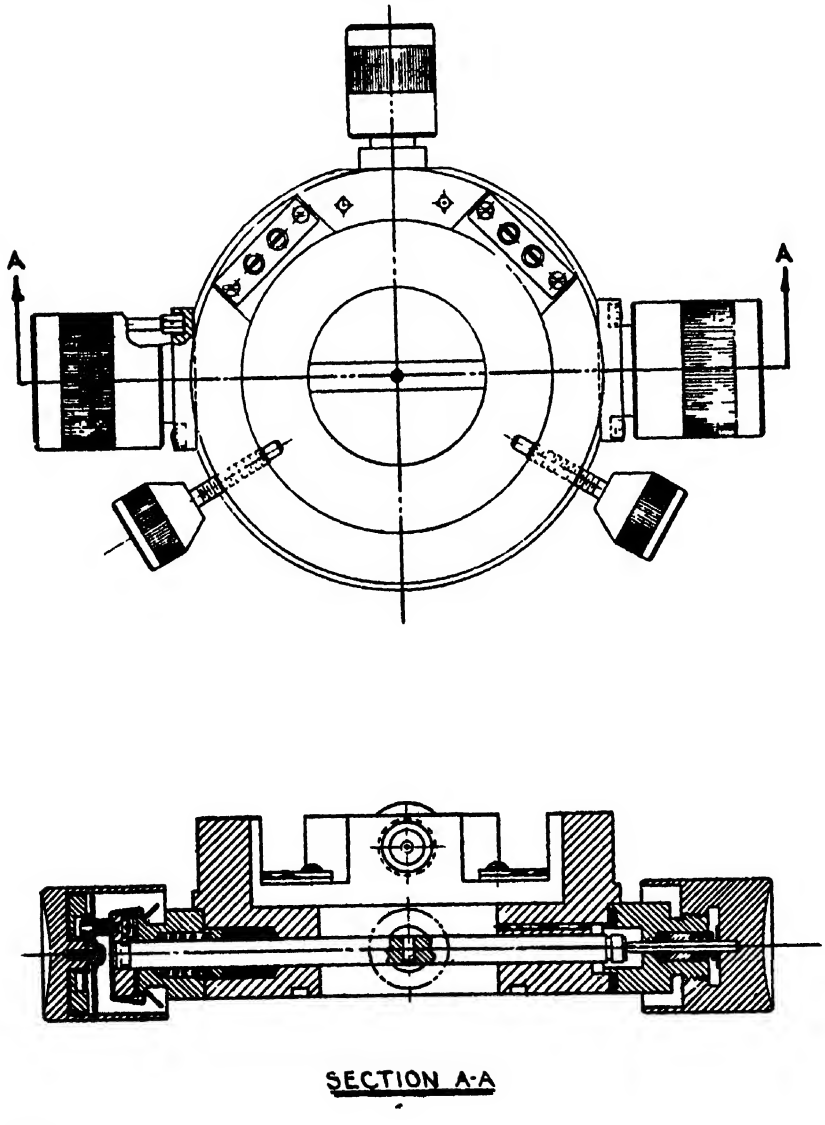


FIG. 4. Cross-sectional view of specimen and electron gun controls.

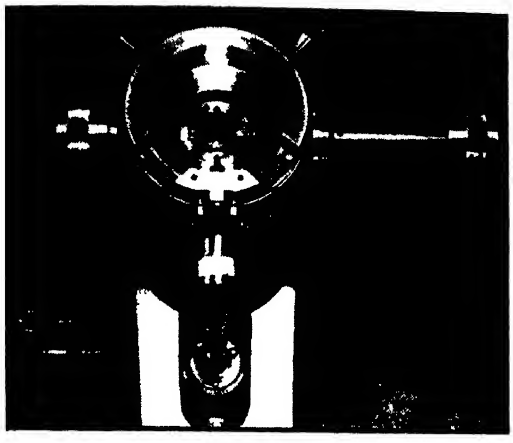


FIG. 5. Rear view of the upper end of the column with the gun assembly removed to show philister head screws used to align the upper element of the objective lens. The large circular body in the center of the chamber is the holder for the "condenser" aperture. The two knobs at the top of the picture adjust the position of the gun assembly.

tamination occurs in all objectives, but negligibly in projector lenses where the beam intensity is very low and the pole faces are completely shielded from the gun. The objective lens of the described microscope is easily demounted for cleaning and replacing apertures. The upper element may be removed through the specimen chamber, at which time the lower element of the objective pole piece is exposed for inspection and minor cleaning. Occasionally, it is necessary to remove the lower objective element. To do this, the projector pole piece is screwed out of its adjustment threads and the assembly pushed down and out of the lens energizing spool. The plate door and shutter should be open to admit the fingers to guide the pole piece out of the viewing chamber. Since the projector pole piece does not contaminate to any appreciable extent, its two pole piece elements are forced permanently onto its nonmagnetic spacer. Large holes in the spacer wall permit access to the pole faces for blowing out dust and for permitting minor cleaning.

The lower objective pole piece element is held tightly to the projector pole piece by set screws which pull down on a conical surface in such a way that the two iron pieces register along an inside surface. Such close registry minimizes stray fringe flux which might influence the electron beam.

Apertures are used (1) to restrict the beam striking the specimen, (2) to intercept electrons scattered from the specimen and thus increase image contrast, and (3) to prevent unused electrons from the powerful projector lens from reflecting off the viewing chamber walls onto the viewing and photographing area to cause "fog." The first two functions are achieved with accurately centered 75 micron platinum apertures. The apertures are identical and may be shifted from the objective to the collimating position as they become too contaminated for use in the objective gap. The small size of the aperture hole protects the specimen from damage from the intense electron beam and from radiant heat and evaporation products from the hot filament. The hole size is large enough to assure alignment of apertures without auxiliary mechanical positioning means and yet is small enough to give good contrast in the objective. The aperture following the projector is a 0.145 inch hole and is not critical in position or dimension.

ALIGNMENT

Alignment of the upper objective pole piece is a problem in vectors. Since the objective misalignment is manifested by an essentially linear shift of the image as the accelerating voltage is varied, the degree of misalignment may be measured by comparing shifts occurring for unit changes of voltage. It is found that the direction of image shift is different from the direction in which the center of the upper objective pole piece deviates from the true optical axis of the microscope. At 6000×, for medium misalignments, corresponding to less than 0.5 inch/1000 volts, the directions differ by about 120°, the misalignment direction being counterclockwise from the image shift when looking along the axis of the instrument toward the viewing screen. The magnitude of the change in image shift per kilovolt change in accelerating voltage is about 0.5 inch/1$K.V.$ for an 0.001 inch misalignment, or 75° turn on the yoke adjustment screws. Once an image shift has been noted as to direction and magnitude, the pole piece many be adjusted a compensating amount

FIG. 6. A side view of the microscope with back cover removed to show arrangement of the vacuum system.

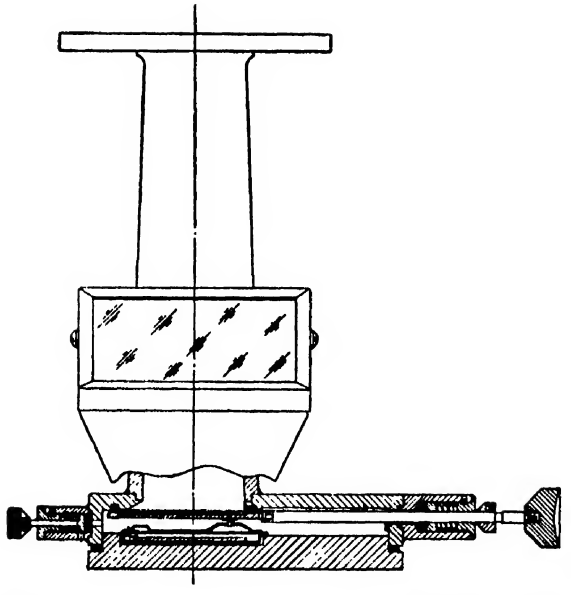


FIG. 7. Cross section showing the sliding value of the photographic chamber. The knob on the right is the activating means. The knob on the left admits air to the photographic chamber to permit opening the chamber door.

in the opposite direction 120° counterclockwise from that of the image shift. Usually, two or three attempts will reduce the image shift to a negligible value so that the only image change with voltage change is that of magnification. As previously stated, once the alignment yoke is adjusted and tightened, the alignment will hold while pole pieces are removed and changed through very long periods of use.

LENS ENERGIZATION

The lens pole piece gaps are energized by permanent magnets appropriately disposed in a magnetic circuit to reduce the deleterious effects of stray fields.* The magnetic means is Alnico V in four rectangular magnets placed parallel to the optical axis of the microscope. The two gaps are energized in parallel, but in the opposite sense, so that the outside shell of the magnetic assembly is sensibly a surface of constant mmf. This largely eliminates stray fields which can reduce magnification, produce asymmetries and distortions in the images, and make alignment difficult. The upper, objective gap gives a magnetomotive force of 1300 gilberts, while the lower projector gap gives a magnetomotive force of 900 gilberts. These values are comparable with those from electromagnetic energizers of the same size. The magnetic circuit parameters are designed so as to give high flux leakage; i.e., to operate the magnets at a large demagnetization coefficient. Such design makes possible the frequent insertion and removal of pole pieces without fear of demagnetization of the energizing magnets. The magnets are energized after assembly. If the spool assembly is ever disassembled, the magnets will be demagnetized to the point where the lenses are not adequately energized.

In spite of the use of powerful magnets, it is not difficult to insert or remove pole piece elements from the bore of the magnetic spool assembly.

VACUUM SYSTEM

The volume of the entire microscope, excluding pumping elements, is approximately one liter. This volume is divided into two pumping circuits. The plate chamber and viewing chamber comprise about 70 percent of the evacuated space, and are pumped directly into the diffusion pump through a $1\frac{3}{8}$ inch opening. The specimen chamber and gun are pumped through a half-inch line disposed along the back of the instrument. The plate chamber accounts for eight percent, the gun 25 percent, and the specimen chamber three percent of the total volume. Figure 6 shows the vacuum system.

Two specialized valving means are provided to replace the versatile, but complex valve systems generally used on electron microscopes. Basically, only two reasons for breaking the vacuum exist in routine electron microscopy; the insertion or removal of specimens, and the insertion and removal of photographic plates. Simple, multi-purpose assembles have been devised to fulfill these specialized functions. The vacuum sealing surface of the specimen lock is also the bearing through which all specimen motion is transmitted into the chamber, while the valve plate which seals the photographic chamber also provides a fluorescent screen and serves as a photographic shutter.

Insertion of a specimen admits only 0.02 cc of air into the vacuum system, while withdrawal of a specimen admits no air. The pumping system will recover safe operating vacuum within six seconds after the admission of a specimen. The introduction of a photographic plate into the vacuum enclosure introduces an appreciable amount of air, 80 cc. This is a result of the necessity for providing space for a sliding valve plate and cassette cover during photographic exposures. Figure 7 shows the valving mechanism. The valve is a sliding door which is moved on and off a circular gasket cemented in the wall of the chamber casting. To pump a plate, the knob outside the chamber is pulled to slide the valve off its gasket. This action breaks the vacuum seal between the viewing chamber and the photographic chamber, permitting the latter to be pumped out. Returning the knob to its initial position seals off the photo-chamber. Air is admitted to the photo-chamber so that it may be opened by pulling a simple relief valve on the left side of the photo-chamber. The vacuum valve system is very dependable. Air may be admited to the photochamber after plate exposure, without turning off the electron beam. A time of about 60 seconds is required to recover operating vacuum after the introduction of a dry photographic plate.

All air introduced with specimens and plates must

* J. H. Reisner, J. App. Phys. 21, 68 (1950). (In process of publication in full.)

pass through the diffusion pump. In spite of the deleterious effect which air has upon diffusion pump oil, octoil has been used for several months of general operation. With the much more stable silicone oil, very much longer periods of operation are possible.

When, in occasional maintenance, any of the components of the microscope are to be adjusted or removed from the vacuum, it is necessary to turn off the diffusion pump for several minutes before admitting air to the system through the relief valve on the photochamber.

The diffusion pump is a small air cooled D.P.I. VMF-10, using a Pressovac as a fore pump. The pump-down time from a cold start is about twenty-five minutes. A thermocouple gauge is used to indicate the vacuum. The meter employed is the test meter of the power supply which may be switched in when a pressure indication is desired.

SPECIMEN CONTROL

An unusual treatment of the specimen introduction and control has resulted in very great simplification in the operation and construction of the microscope. Grommet vacuum seals about shafts have long been used to obtain displacement and rotation through the confines of vacuum systems. By putting small cut-outs in such shafts, it is possible to introduce objects placed in the cut-out space into the vacuum, if the grommet is longer than the cut-out space. Such is the basis for the specimen holder to be described. Figures 4 and 5 show the actual specimen control system. A cylindrical hole 0.140 diameter, to accommodate the specimen holder, is open perpendicularly through a $\frac{5}{16}$ diameter shaft to bottom 0.012 inch from the surface of the shaft. A $\frac{1}{16}$ hole is continued through the shaft wall to permit egress of the electron beam. The transverse hole is roughly at the center of the six inch shaft. It is so positioned that when the shaft is pushed into the instrument up to the plunger of the right hand control, the specimen holder will be positioned over the lens opening. The area of the specimen is scanned in one direction by driving the shaft axially with the right-hand knob and plunger, and in a perpendicular direction by rotating with the left-hand knob.

The shaft rotation amounts to but 3° from its center position and permits scanning of three openings of a 200 mesh specimen support screen. The screen is situated 0.139 inch from the axis of rotation of the specimen shaft. In the extreme rotation positions, the specimen is only five microns closer to the objective lens than in its mean position. Such a displacement represents about 70 volts difference in focal voltages with the 3000× pole piece, and about 200 volts with the 6000× pole piece. The depth of focus is so great with the electron microscope that it is generally not necessary to refocus the field in going from the center to the extremes of specimen holder rotation. A total of about 36 mesh openings may be observed, 12 along the axis of the rod, three by rotation.

The axial movement of the specimen rod is provided through the right-hand knob by means of screw threads which advance a plunger registering against a hard metal plate on the end of the specimen holder rod. The rod is held tightly against the plunger by springs in the left-hand knob which grip a conical surface on the microscope housing. A limit spring slips into a groove at the right end of the specimen holder rod to prevent the rod from being pulled out of its sealing grommet and to prevent it from sliding back into the instrument under atmospheric pressure. The ability to remove the holder entirely from the path of the beam is useful in initially lining up the gun with the other optical elements.

Because of the sensitivity of the transverse control, six degrees of rotation to give full range of motion, the left-hand control knob is geared down 1:3. It has been found from experience that under these circumstances, control is quite adequate. The left-hand knob is highly damped in action to resist twist in the vacuum sealing grommet. In order that the specimen rod be correctly oriented to pass the electron beam through the specimen, a guide pin on the microscope must fit into a slot on the left hand knob before the rod can engage its springs on the advancing cone.

In order to keep the vacuum sealing grommet from wearing or binding, the specimen holder shaft is very lightly lubricated periodically with clear octoil.

VIEWING CHAMBER ASSEMBLY

In all electron microscopes, the viewing screen and plate must be normal to the beam. Direct viewing, desirable for brightness and sharpness, means that the fluorescent screen must be viewed obliquely. Therefore, the fifteen degree slope of this microscope column has been applied to facilitate direct viewing of the image. A wide window permits viewing with both eyes, and viewing by two people simultaneously.

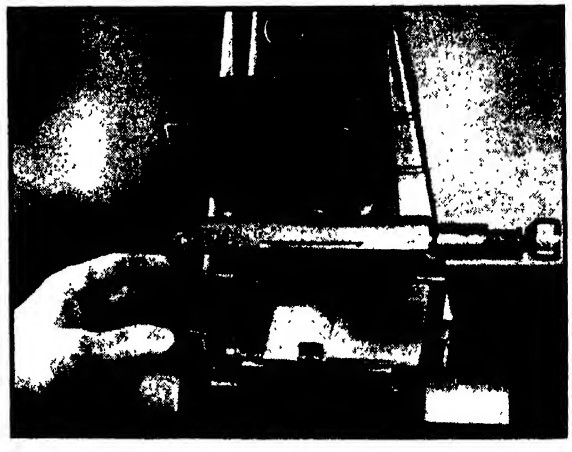


FIG. 8. Insertion of a photographic plate cassette.

Valving and plate exposure are accomplished by two moving parts, a rod with a hand knob on the end, and a metal plate. The plate has a circular depressed area on its upper surface to hold the electron sensitive phosphor and yet protect it from the vacuum sealing gasket across which the metal plate slides in the various phases of its valving cycle. A pin on the lower surface of the metal plate engages the cover to the photographic plate cassette. The description of a complete photographic cycle will make the operation of the mechanism more apparent.

With the camera door dropped open, a 2×2 plate cassette is clipped into an orienting slot and hold-down springs, Fig. 8. The door is snapped shut. The photochamber is pumped by sliding the movable plate far enough to the right to slip off the sealing gasket and open the chamber to the column. A stop on the shaft prevents more than the correct displacement. As soon as the photo chamber is evacuated, the high voltage may be turned on and an image focused. To expose the photographic plate, the same right hand knob is pulled to its extreme right position. This withdraws the valve plate from before the photographic plate. At the same time, a catch on the valve plate has engaged the cassette cover and slid it to the side, thus exposing the photographic plate. To end the exposure, the knob is pushed back to its inmost position, which closes the cassette cover and returns the valve plate to cover the vacuum gasket. A pull on a release valve on the left side of the photochamber admits air to the chamber which may then be opened for cassette removal.

The sliding valve plate is held against the vacuum gasket by light spring runners. The opening in the base of the viewing chamber is $1\frac{7}{8}$ inches in diameter, permitting exposure of a circular area of the same diameter on the photographic plate. The pole piece may be inserted and withdrawn through this hole with the aid of simple tools. The objective gap may be varied by engaging the screw slots on the projector by means of another simple tool inserted through the bottom hole when the valve plate is pulled aside.

POWER SUPPLY

Accelerating voltage and filament heating current are supplied by a fifty kilovolt supply almost identical with that used on commercial microscopes.[3] A driver unit which can sit on the table near the microscope permits control of voltage for focusing and control of emission for beam intensity. Once set with the variable controls, the chosen voltage and filament current reappear with every turn-on of the high voltage power switch. A meter reads, over its full scale, the reference voltage corresponding to the output range of voltages from 46 to 50 kv. Thus the focus voltage may be observed accurately on the meter. Such a feature is an excellent check on nearness to focus. It is useful in making through-focus series of photographs. It is a necessity in adjusting pole pieces to attain the desired focal voltage.

All electrical controls cause a voltage increase with a clockwise rotation. This means that the focal length of the lenses increases with a clockwise rotation of the high voltage focus control. As a result of the same electrical convention, electromagnetic lenses have their focal length decreased by the clockwise rotation. Using voltage focusing, the experienced electronmicroscopist finds the changes of appearance of fringes in the image occurring in reverse order to that which is encountered with current focusing. No other difference is apparent in the focusing process.

In an instrument designed for use by personnel untrained in electrical practice, safety precautions must be extensive. Circuit interlocks necessitate the correct positioning of the high voltage cable on the instrument and the attachment of the protective gun hood before voltage can be applied to the instrument. A quick-acting grounding mechanism is provided to discharge the high voltage supply when the gun hood is raised from its seat more than $\frac{3}{8}$ of an inch. If the high voltage has not been turned off before raising the hood, grounding the output in this manner will cause an overload relay system on the power supply to turn off the power.

A lead glass viewing window is provided to prevent the escape of x-rays generated by the electrons incident upon the viewing screen.

GENERAL FEATURES

The electron gun is the only externally adjustable optical element of the electron microscope. A heavy brass anode structure supports the filament grid and glass insulator components which are rigidly affixed to it. The brass anode is also the movable element, permitting sliding movement in a plane perpendicular to the optical axis of the microscope. The flat lower wall of the anode structure has holes to pass the electron beam and to permit evacuation of the gun volume through the specimen chamber. A gasket seat is provided on the bottom surface to carry the rubber gun gasket which provides a sliding vacuum seal between the anode and the specimen chamber. The position of the gun axis is determined by two adjustable plungers which register radially against the cylindrical outside wall of the anode. These plungers are activated by knobs, the two highest knobs on the column (Fig. 5), working through screw threads. A spring-loaded plunger extending radially backward on the microscope holds the anode tightly against the adjusting plungers to make the anode follow the plungers. The complete gun assembly may be removed from the column by removing the spring loaded plunger and sliding the gun to its extreme rear position so as to clear the adjusting plungers. If the position of the gun adjustment knobs is not changed, a gun may be removed and reinserted without changing alignment enough to lose illumination.

Since filaments, apertures, and diffusion pump oil

must be replaced occasionally and pole pieces must be removed for cleaning, all these components may be reached without major disassembly of the instrument or removal from its stand.

Stray alternating magnetic fields and vibration have always been two serious environmental hazards in high resolution electron microscopy. Careful consideration has been given toward overcoming these dangers in this instrument. From the objective pole piece to the photographic plate, the microscope beam is shielded behind heavy iron walls. In the extremely sensitive field region between lenses, the shielding from external fields is complete. Operation in the vicinity of relatively strong fields has shown little effect on the microscope image.

Vibration problems are largely eliminated by rigid construction throughout. Lens elements are held fast by the magnetic field in a one-piece housing. The viewing and specimen chambers are rigidly bolted to the lens housing. The two moving parts are the specimen rod and gun, both strong and well supported. The column of the instrument is supported at the base of the lens housing, not far from the center of gravity of the column, with the result that environmental vibration acts on the whole rather than on various parts.

SUMMARY

The instrument herein described has shown itself to be excellent for almost all types of electron microscopy not requiring resolutions better than 100A. Its unusual reproducibility makes it ideal for routine microscopy. Long sequences of pictures may be taken without touching gun adjustment, focus, or emission controls.

The direct magnifications of 1500, 3000, and 6000 times are chosen for convenience of observation and are not to be taken as a measure of resolution. Well-focused photographic images frequently reveal useful detail up to 20,000 diameters over-all magnification. The 1500× lens system is useful where studies are to be made on specimens familiar to the light microscopist and on specimens such as sections of tissue where fine structure must be viewed in relation to familiar gross structure such as cell walls. The 6000× magnification is useful where the highest resolution is routinely desired, such as in virus studies, and may most easily be achieved by virtue of the ability to focus the high magnification optical image most accurately. Figure 9 shows a high resolution photomicrograph taken with the described instrument.

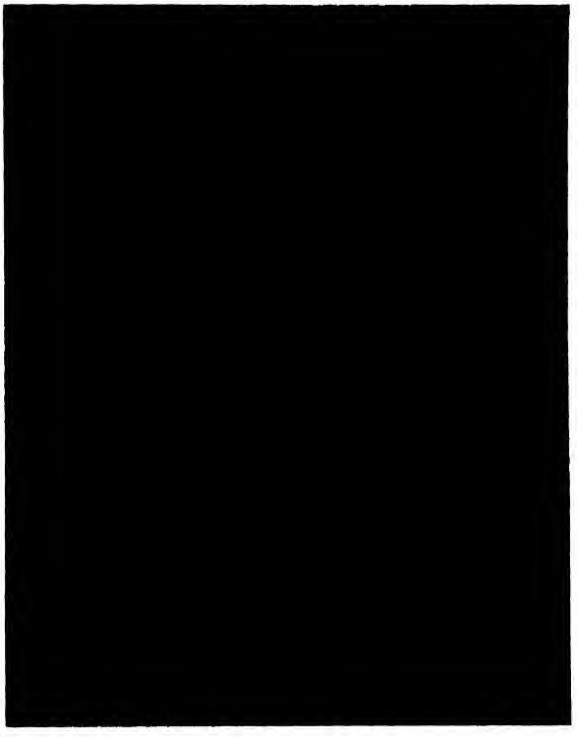


FIG. 9. Photomicrograph of a shadowed replica of an etched stainless steel surface. Magnification 20,000×. Taken at 3000×.

ACKNOWLEDGMENT

The authors wish to acknowledge the guidance and technical assistance of Perry C. Smith, under whose direction the work was carried out, and of Robert G. Picard, who frequently collaborated on the project. The electrical circuitry was carried out by Samuel M. Zollers. Acknowledgment is due Charles Yearsley, Richard McGarry, and A. M. Rennie for their assistance on mechanical design, and to Gustave Burger for much of the actual construction.

Artificial Field Equations for a Region Where μ and ϵ Vary with Position

P. D. P. SMITH
Westmount, Province of Quebec, Canada
(Received March 6, 1950)

I. Theory. The field equations developed admit of solutions in regions where μ and ϵ vary with position. For regions of zero conductivity, they are symmetrical with respect to electric and magnetic quantities, and the TM case can be derived from the TE. Fields derivable from a single component 4-potential are examined, and their propagation equations are found. A table relates the tensor and pseudo-tensor components to the usual vector components.

II. Applications in cylindrical polar coordinates. For the TE case, three types of solution exist, called A, B, or C, where the **E** field is parallel to the elementary vectors **dz**, **dr** or **rdθ**, respectively (for the TM case, **H** replaces **E**). In each TM or TE case, two possible distributions of μ and ϵ occur, leading to separable propagation equations. TEM waves can be derived from both TM and TE cases, but the impedances are different.

TE, TM and TEM waves for certain specific distributions of μ and ϵ, and their applications to curved guides, horns, and co-axial lines are considered. Some expressions are obtained for the fields, propagation constants, cut-off frequencies, and impedance. These solutions are not necessarily of the most general type, for to meet the conditions that restrict their choice, μ and ϵ are taken to be independent of the coordinate to which the transverse component of **E** (in TE) or **H** (in TM) is parallel.

I. THEORY

THE usual methods of procedure yield solutions of Maxwell's equations only in the following cases: (a) where space is everywhere homogeneous, (b) where space is subdivided into several homogeneous regions, (c) where the inductivity varies so slowly that it may be considered homogeneous.

This has been pointed out by Liebowitz,[1] who showed that it is possible to derive separable wave equations in a space where ϵ* is a function of position, by making use of a metric

$$ds^2=1/\epsilon\cdot\{(du^1)^2+(du^2)^2+(du^3)^2\}$$

and an auxiliary scalar function which is determined from ϵ. He proceeded to investigate the static field of a point charge. By assuming suitable distributions of ϵ, the field energy could be rendered finite.

However, a statement of the distribution of ϵ is a statement of the variation of the speed of light, and therefore of the variation of gravitational potential. It appears that his attack is really equivalent to guessing the solutions of Einstein's field equations for the space around a point charge. This particular instance does not concern the reduction in light velocity by scattering from many centers, as occurs in a dielectric, but a reduction in light velocity supposed inherent in the properties of space itself, such as that which causes the deviation of a light ray passing a heavy body. The problem of the gravitational field of a point charge has been solved by Weyl[2] and others, so that it is not necessary to guess at the result. The more fruitful use of equations of this type may be in problems of applied physics, to find solutions of Maxwell's equations for regions where μ and ϵ are known functions of position.

An alternative approach to Liebowitz's problem is formulated in this paper. The procedure used here is to set up a set of tensor field equations for a space where μ and ϵ are functions of position, and then investigate certain simple types of transverse electric and transverse magnetic waves, and the conditions for their propagation. The equations give results for regions where both μ and ϵ vary with position, and this respect are more general than Liebowitz's. On the other hand, solutions have only been found for certain simple cases where the four potential has only one component.

It turns out to be profitable to seek distributions of μ and ϵ leading to separable equations of propagation, rather than to seek general solutions for arbitrary μ and ϵ.

The field equations used are "artificial," in the sense that they are an artifice to help obtain solutions for the case of variable μ and ϵ. They have the appearance of being relativistic, but actually are only exact when the medium is at rest with respect to the observer. There is no practical disadvantage in this from an engineering point of view, as the velocity of the medium with respect to the observer is always far smaller than the velocity of light in the medium.

The equations, which have been published previously,[3] are

$$F^i=\partial_j F^{ij}, \tag{1.01}$$

$$0=\partial_i\varphi_{jk}+\partial_j\varphi_{ki}+\partial_k\varphi_{ij}, \tag{1.02}$$

$$F^{ij}=g^{\frac{1}{2}}\cdot(\epsilon/\mu)^{\frac{1}{2}}\varphi^{ij}, \tag{1.03}$$

$$F^i=g^{\frac{1}{2}}\rho\cdot dx^i/ds, \tag{1.04}$$

$$ds^2=g_{ij}dx^i dx^j=c^2dt^2/\mu\epsilon-\alpha^2(dx^1)^2-\beta^2(dx^2)^2-\gamma^2(dx^3)^2\simeq c^2dt^2/\mu\epsilon. \tag{1.05}$$

Equation (1.03) differs from the usual form because of the factor $(\epsilon/\mu)^{\frac{1}{2}}$. The current tensor, (1.04), becomes in an orthogonal coordinate system

$$F^4=\alpha\beta\gamma\rho\dot{x}^4/c(1-v^2/v_0^2)^{\frac{1}{2}},$$
$$v_0=c/(\mu\epsilon)^{\frac{1}{2}}.$$

[1] Liebowitz, Phys. Rev. 64, 294 (1943).
* ϵ, the dielectric "constant."
[2] Weyl, *Space, Time, Matter* (Methuen and Company, Ltd., London, 1920), pp. 252–262.
[3] P. D. P. Smith, J. App. Phys. 20, 633 (1949).

This shows that the expressions are only accurate when $v \ll v_0$, and are not truly relativistic.

Ohm's law, for this system, becomes

$$f_i = \sigma \varphi_{a,i} \bar{u}^a, \quad i<4, \tag{1.06}$$

where $\bar{u}^a$ is the mean velocity defined by

$$\bar{u}^a = \int u^a \cdot dN \Big/ \int dN,$$

$\int dN$ being the number of electrons in a volume that is small enough so that g_{ij} may be considered a constant, σ representing the conductivity. The Eq. (1.01) becomes

$$\{\sigma \cdot (\mu/\epsilon)^{\frac{1}{2}} \bar{u}_j + \partial_j\} F^{ij} = 0, \quad i<4 \tag{1.07}$$

equivalent to

$$\sigma\{\mathbf{E} + (1/c)[\bar{\mathbf{v}} \times \mathbf{B}]\} = \text{curl}\mathbf{H} - \dot{\mathbf{D}}/c \tag{1.08}$$

in vector notation.

In regions where $\sigma = 0$, there is complete symmetry between the electric and magnetic fields when the roles of μ and ϵ are interchanged. This symmetry can be brought out by introducing a new four-potential, ψ_i, the counterpart of the usual φ_i. We had

$$\varphi_{ij} = \partial_i \varphi_j - \partial_j \varphi_i \tag{1.09}$$

from which can be defined the pseudo-tensor

$$B^{ij} = (-)^{p(i,j,k,l)} \varphi_{kl}, \tag{1.10}$$

where $p(i, j, k, l)$ is the number of inversions of i, j, k, l from the order 1, 2, 3, 4. Thus it follows

$$F^{ij} = (g\epsilon/\mu) g^{ia} g^{jb} (-)^{p(a,b,c,d)} B^{cd} \tag{1.11}$$

and conversely

$$B^{ij} = (\mu/g\epsilon) g_{ka} g_{lb} (-)^{p(i,j,k,l)} F^{ab}. \tag{1.12}$$

At this point the second type of field will be introduced, namely that produced by the alternative four-potential ψ_i. For these fields

$$\psi_{cd} = \partial_c \psi_d - \partial_d \psi_c, \tag{1.13}$$

$$\partial_b \psi_{cd} + \partial_c \psi_{db} + \partial_d \psi_{bc} = 0 \tag{1.14}$$

and

$$F^{ab} = (-)^{p(a,b,c,d)} \psi_{cd}. \tag{1.15}$$

Using this in (1.12),

$$B^{ij} = (-)^{p(i,j,k,l)+p(a,b,c,d)} (\mu/g\epsilon) g_{ka} g_{lb} \psi_{cd}. \tag{1.16}$$

The majority of practical problems will be set up in orthogonal coordinates, in which (1.16) simplifies to

$$B^{ij} = (-)^{p(i,j,k,l)+p(k,l,i,j)} \cdot (\mu/g\epsilon) g_{kk} g_{ll} \psi_{ij}$$
$$= (\mu/g\epsilon) g_{kk} g_{ll} \psi_{ij}. \tag{1.17}$$

Thus in orthogonal coordinate systems

$$B^{ij} = (\mu/\epsilon)^{\frac{1}{2}} (g_{kk} g_{ll}/g_{ii} g_{jj})^{\frac{1}{2}} \psi_{ij} \tag{1.18}$$

TABLE I. The co-variant and contra-variant fields generated from the four potentials φ_i and ψ_i, and their equivalents in vector notation. The components generated by φ_3 or ψ_3 are indicated by an asterisk.

TYPE 1 FIELD 4 POTENTIAL φ_i	TYPE 2 FIELD 4 POTENTIAL ψ_i
$\varphi_{41} = i\alpha E_1$	$\psi_{41} = -\alpha H_1$
$\varphi_{42} = i\beta E_2$	$\psi_{42} = -\beta H_2$
$\varphi_{43} = i\gamma E_3$ *	$\psi_{43} = -\gamma H_3$ *
$F^{41} = i\beta\gamma D_1$	$B^{41} = -\beta\gamma B_1$
$F^{42} = i\gamma\alpha D_2$	$B^{42} = -\gamma\alpha B_2$
$F^{43} = i\alpha\beta D_3$ *	$B^{43} = -\alpha\beta B_3$ *
$F^{12} = \gamma H_3$	$B^{12} = -i\gamma E_3$
$F^{23} = \alpha H_1$ *	$B^{23} = -i\alpha E_1$ *
$F^{31} = \beta H_2$ *	$B^{31} = -i\beta E_2$ *
$\varphi_{12} = \alpha\beta B_3$	$\psi_{12} = -i\alpha\beta D_3$
$\varphi_{23} = \beta\gamma B_1$ *	$\psi_{23} = -i\beta\gamma D_1$ *
$\varphi_{31} = \gamma\alpha B_2$ *	$\psi_{31} = -i\gamma\alpha D_2$ *

and corresponding to this in the first type field

$$F^{ij} = (\epsilon/\mu)^{\frac{1}{2}} (g_{kk} g_{ll}/g_{ii} g_{jj})^{\frac{1}{2}} \varphi_{ij}. \tag{1.19}$$

From (1.18) and (1.19) is built the following table connecting the tensor and pseudo-tensor components with the usual vector components. As before[3] $\varphi_{41} = i\alpha E_1$, $\varphi_{12} = \alpha\beta B_3$, and use is made Eqs. (1.03), (1.10), and (1.15).

Thus corresponding to any solution of the field equations generated by the 4-potential φ_i there can be found another solution generated by the alternative potential ψ_i. This use of two alternative solutions is very common in wave guide theory, one solution yielding transverse electric, or TE, waves, the other transverse magnetic, or TM, waves.

The TE Case

For the sake of simplicity, the examples solved will be restricted to the case where there is only one non-vanishing component of the 4-potential, so that

$$\varphi_i = \delta_{i3} \varphi. \tag{1.20}$$

The field equations give immediately

$$\begin{aligned} &\varphi_{12} = 0, && \varphi_{41} = 0, \\ &\varphi_{23} = \partial_2 \varphi, && \varphi_{42} = 0, \\ &\varphi_{31} = -\partial_1 \varphi, && \varphi_{43} = \partial_4 \varphi. \end{aligned} \tag{1.21}$$

The contravariant field components are, from (1.19) and (1.21),

$$\begin{aligned} &F^{12} = 0, && F^{41} = 0, \\ &F^{23} = \alpha/\beta\gamma\mu \cdot \partial_2 \varphi, && F^{42} = 0, \\ &F^{31} = -\beta/\gamma\alpha\mu \cdot \partial_1 \varphi, && F^{43} = (\epsilon\alpha\beta/\gamma) \partial_4 \varphi. \end{aligned} \tag{1.22}$$

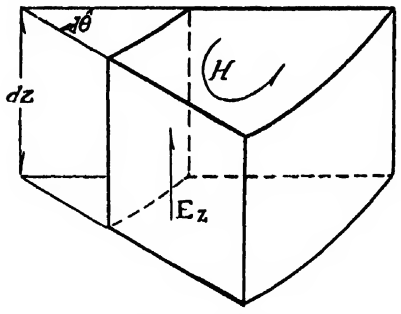


FIG. 1. TE waves of Type A. Fields generated by a single component 4-potential. The electric field is parallel to the z-axis, propagation is radial, or perpendicular to the axial planes.

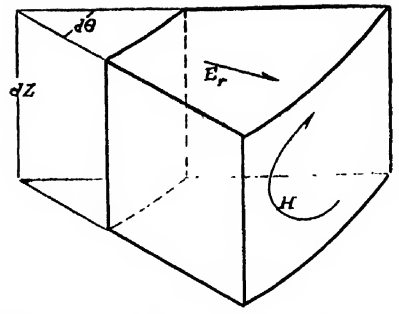


FIG. 2. TE waves of Type B. Fields generated by a single component 4-potential. The electric field is radial, and propagation is along the z-axis, or perpendicular to axial planes.

These non-vanishing components arising from φ_3, ψ_3 are marked with an asterisk in Table I, when their vector counterparts may be read.

Thus the contravariant group of field equations

$$0=\partial_j F^{ij}$$

become

$$0=\partial_1(\beta/\gamma\alpha\mu\cdot\partial_1\varphi)+\partial_2(\alpha/\beta\gamma\mu\cdot\partial_2\varphi)+\partial_4(\epsilon\alpha\beta/\gamma\cdot\partial_4\varphi)=0 \quad (1.23)$$

and also the three equations

$$\partial_3 F^{i3}=0. \quad (1.24)$$

These equations obviously give a TE wave propagated in directions perpendicular to dx^3. The electric field is parallel to the third coordinate dx^3.

If α, β, γ are independent of x^3, the equations (1.24) are identically satisfied if μ, ϵ, and φ are also independent of x^3. This restriction has been used in all cases examined in this paper.

On the other hand, it is simple to eliminate the field quantities from Eqs. (1.24), obtaining a pair of differential equations involving μ and ϵ. These, included merely for completeness, are

$$\partial_3 \log\{\partial_1 X/\partial_3(n^2\alpha^2)\}=0, \quad (1.25)$$

$$\partial_3 \log\{\partial_2 X/\partial_3(n^2\beta^2)\}=0, \quad (1.26)$$

$$n^2=\mu\epsilon, \quad X=\partial_3 \log\{\gamma/\epsilon\alpha\beta\}. \quad (1.27)$$

The TM Case

The TM case is obtained immediately from the TE case by interchanging the roles of μ and ϵ, ψ and φ. There follows for the non-vanishing field components

$$\psi_{23}=\partial_2\psi, \quad \psi_{31}=-\partial_1\psi, \quad \psi_{43}=\partial_4\psi, \quad B^{12}=\alpha/\beta\gamma\epsilon\cdot\partial_2\psi, \quad (1.28)$$

$$B^{21}=-\beta/\gamma\alpha\epsilon\cdot\partial_1\psi, \quad B^{43}=\mu\alpha\beta/\gamma\cdot\partial_4\psi. \quad (1.29)$$

The propagation equation becomes

$$0=\partial_1(\beta/\gamma\alpha\epsilon\cdot\partial_1\psi)+\partial_2(\alpha/\beta\gamma\epsilon\cdot\partial_2\psi)+\partial_4(\mu\alpha\beta/\gamma\cdot\partial_4\psi) \quad (1.30)$$

and also

$$\partial_3 B^{12}=0. \quad (1.31)$$

As before, 1.31 are satisfied if α, β, γ, μ, ϵ, and ψ are independent of x^3. If this is not so, then it would be necessary to ensure that μ and ϵ satisfy

$$\partial_3 \log\{\partial_1 Y/\partial_3(n^2\alpha^2)\}=0,$$
$$\partial_3 \log\{\partial_2 Y/\partial_3(n^2\beta^2)\}=0,$$

where

$$Y=\partial_3 \log(\gamma/\mu\alpha\beta).$$

This completes the theoretical part of this work. The next part involves the application of the theory to certain problems in cylindrical polar coordinates.

II. SOLUTIONS IN CYLINDRICAL POLARS

1. General Properties in Cylindrical Coordinates

Before proceding to consider specific examples using the theory developed in Section I, the general nature of the solutions which can arise from a single component of potential will be considered. The TM case has to be restricted to coordinate systems in which x^1 constant, x^2 constant, is a closed curve, because there is only one B component, and the corresponding lines of flux must be closed curves.

The possible choices are shown in Table II and Table III, along with the identifications of dx^1, dx^2, dx^3, α, β, and γ, and the corresponding propagation equation. The nature of the fields is shown diagrammatically in Figs. 1, 2, and 3 for TE waves, Fig. 4 for TM waves.

The fields are classified as Type A, B, or C, according as the coordinates (dx^1, dx^2, dx^3) are taken as (dr, $d\theta$, dz), ($d\theta$, dz, dr) or (dz, dr, $d\theta$).

The non-vanishing field components can be verified by referring to Table I of Section I, and remembering that only the third component of potential, φ_3 or ψ_3, is to be considered.

The form of TM waves conjugate to the TE waves of type C are shown in Table III. The equations $\partial_3 B^{12}$ are automatically satisfied if ψ, μ, ϵ are independent of θ, since α, β, and γ are independent of θ.

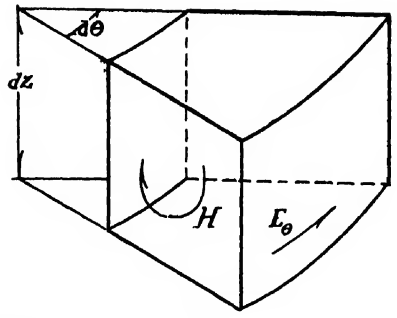


FIG. 3. TE waves of Type C. Fields generated by single component 4-potential. The electric field is perpendicular to planes through the axis, and propagation is radial or along the axis.

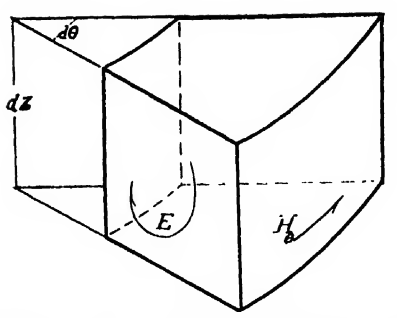


FIG. 4. TM waves of Type C. Fields generated by a single component of the alternative 4-potential. The magnetic field is perpendicular to planes through the axis, and propagation is radial, or along the axis.

TEM Waves

In addition to TE and TM waves, TEM waves can be propagated also. These are characterized by the potential depending on only one coordinate variable. For example, in the TM case the electric fields are given by

$$-i\beta\gamma\cdot D_1=\partial_2\psi, \quad -i\gamma\alpha\cdot D_2=-\partial_1\psi. \tag{2.01}$$

For TEM waves in a co-axial line, $D_z=0$, and hence it is necessary that $\partial\psi/\partial r=0$. The propagation equation reduces to

$$\frac{\partial}{\partial z}\left\{1/\epsilon\cdot\frac{\partial\psi}{\partial z}\right\}+\omega^2/c^2\mu\psi=0. \tag{2.02}$$

The field components are then given by

$$D_r=-i/r\cdot\partial\psi/\partial z, \quad H_\theta=-\omega/cr\cdot\psi. \tag{2.03}$$

The ratio

$$Z=E_r/H_\theta=ic/\omega\epsilon\cdot\partial/\partial z\log\psi. \tag{2.04}$$

The TE fields of type B, with $B_z=0$, can also yield a TEM wave. If $B_z=0$, then referring to the propagation equation for type B waves on Table II, in this case, where

$$B_z=-1/r\cdot\partial\varphi/\partial\theta=0, \tag{2.05}$$

it becomes

$$\partial/\partial z\{1/\mu\cdot\partial/\partial z(r\varphi)\}+\omega^2/c^2\cdot\epsilon\cdot r\varphi=0. \tag{2.06}$$

The field impedance ratio in this case becomes

$$Z'=E_r/H_\theta=(\omega\mu/ic)/\partial/\partial z\log\varphi. \tag{2.07}$$

If ψ and φ are proportional to $e^{S \text{ or } T}$, then for the TEM waves derived from TM waves

$$Z=\{\mu/\epsilon+c^2/\omega^2\epsilon^2\cdot(S''-S'\epsilon'/\epsilon)\}^{\frac{1}{2}}, \tag{2.08(a)}$$

while for those derived from TE waves

$$Z'=\{\epsilon/\mu+c^2/\omega^2\mu^2(T''-T'\mu'/\mu)\}^{-\frac{1}{2}}. \tag{2.08(b)}$$

When μ and ϵ are constant,

$$Z=Z'.$$

The next sections deal with specific examples of the method in greater detail. Only a few of the many possible examples have been solved.

Section 2. TE Waves of Type A

In this case

$$dx'=dr, \quad dx^2=d\theta, \quad dx^3=dz, \quad \alpha=1, \quad \beta=r, \quad \gamma=1. \tag{2.09}$$

The non-vanishing field components are

$$E_z=\omega/ic\cdot\varphi,$$
$$B_r=1/r\cdot\partial\varphi/\partial\theta, \quad B_\theta=-\partial\varphi/\partial r. \tag{2.10}$$

The propagation equation becomes, in this case

$$0=\frac{\partial}{\partial r}\left\{\frac{r}{\mu}\frac{\partial\varphi}{\partial r}\right\}+\frac{\partial}{\partial\theta}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial\theta}\right\}+\frac{\omega^2}{c^2}\epsilon r\varphi. \tag{2.11}$$

The remaining equations,

$$\partial/\partial z(F^{i3})=0 \tag{2.12}$$

are automatically satisfied if φ, μ, and ϵ are independent of z, because α, β, γ are all independent of z.

The next problem is to obtain a possible pair of functions

$$\mu=\mu(r,\theta), \quad \epsilon=\epsilon(r,\theta)$$

that make (2.11) separable. Let

$$\mu=a(r)b(\theta), \quad \epsilon=c(r)f(\theta) \tag{2.13}$$

and let

$$\varphi=R(r)\Theta(\theta).$$

Then Eq. (2.11) becomes

$$\frac{ra(r)}{R}\frac{d}{dr}\left\{\frac{r}{a(r)}\frac{dR}{dr}\right\}+\frac{b(\theta)}{\Theta}\frac{d}{d\theta}\left\{\frac{1}{b(\theta)}\frac{d\Theta}{d\theta}\right\}$$
$$+\omega^2/c^2\cdot r^2a(r)c(r)b(\theta)f(\theta)=0.$$

This equation is separable if either

$$r^2a(r)c(r)=k_r \text{ a constant, or,} \tag{2.14}$$

$$b(\theta)f(\theta)=k_\theta \text{ a constant.} \tag{2.15}$$

These two cases will be referred to as k_r-separability and k_θ-separability.

TABLE II. Propagation equations and field components for TE waves generated by a single component of 4-potential.

dx^1	dx^2	dx^3	α	β	γ	Propagation Equation + Field Components	Character of Waves
dr	$d\theta$	dz	1	r	1	$\frac{\partial}{\partial r}\left\{\frac{r}{\mu}\frac{\partial\varphi}{\partial r}\right\}+\frac{\partial}{\partial\theta}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial\theta}\right\}+\frac{\omega^2}{c^2}\epsilon r\varphi=0$ $E_z=\omega/ic\,\varphi,\qquad B_r=1/r\cdot\frac{\partial\varphi}{\partial\theta}$ $B_\theta=-\frac{\partial\varphi}{\partial r}$	(a) TE Waves in a curved guide travelling in direction $d\theta$. (b) TE Waves in a horn, travelling in direction dr. (c) TE Waves travelling radially in a cavity with cylindrical walls.
	TYPE A						
$d\theta$	dz	dr	r	1	1	$\frac{\partial}{\partial\theta}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial\theta}\right\}+\frac{\partial}{\partial z}\left\{\frac{r}{\mu}\frac{\partial\varphi}{\partial z}\right\}+\frac{\omega^2}{c^2}\epsilon r\varphi=0$ $E_r=\omega/ic\,\varphi\qquad B_\theta=\frac{\partial\varphi}{\partial z}$ $B_z=-1/r\frac{\partial\varphi}{\partial\theta}$	(a) TE Waves in a guide with inclined walls, travelling in direction dz. (b) TE Waves in a curved guide, travelling in direction $d\theta$
	TYPE B						
dz	dr	$d\theta$	1	1	r	$\frac{\partial}{\partial z}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial z}\right\}+\frac{\partial}{\partial r}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial r}\right\}+\frac{\omega^2\epsilon}{c^2}\varphi=0$ $E_\theta=\omega/icr\cdot\varphi\qquad B_z=1/r\cdot\frac{\partial\varphi}{\partial r}$ $B_r=-1/r\cdot\frac{\partial\varphi}{\partial z}$	(a) TE Waves in a guide with inclined walls travelling in a direction dz. (b) TE Waves in a horn, travelling in a direction dr.
	TYPE C						

For k_r-separability

$$\frac{ra(r)}{R}\frac{d}{dr}\left\{\frac{r}{a(r)}\frac{dR}{dr}\right\}=\lambda^2, \tag{2.16}$$

$$0=\frac{b(\theta)}{\Theta}\frac{d}{d\theta}\left\{\frac{1}{b(\theta)}\frac{d\Theta}{d\theta}\right\}+\{\lambda^2+\omega^2/c^2\cdot k_r b(\theta)f(\theta)\}. \tag{2.17}$$

On the other hand, k_θ-separability gives

$$\frac{ra(r)}{R}\frac{d}{dr}\left\{\frac{r}{a(r)}\frac{dR}{dr}\right\}=\{\eta^2-\omega^2/c^2\cdot k_\theta\cdot r^2a(r)c(r)\}, \tag{2.18}$$

$$\frac{b(\theta)}{\Theta}\frac{d}{d\theta}\left\{\frac{1}{b(\theta)}\frac{d\Theta}{d\theta}\right\}+\eta^2=0. \tag{2.19}$$

An Example of TE Waves of Type A Using k_θ-Separability

Either of these last two alternatives can be attacked as a new differential equation, the values of μ and ϵ being given. An alternative attack is to choose the values of $a(r)$ and $c(r)$ in such a way as to generate known differential equations, or, what is equivalent, to choose the functions R and Θ and seek as solutions $a(r)$ and $c(r)$. This provides a great flexibility for design of apparatus. The function $b(\theta)$ must be periodic if θ can vary over a range greater than 2π; for example, it could be

$$b(\theta)=1+A\cos k\theta$$

or it could even be represented by a Fourier series.

Equation (2.19) can also be written

$$d^2\Theta/d\theta^2+d\Theta/d\theta\cdot d/d\theta(\log 1/b(\theta))+\eta^2\Theta=0. \tag{2.20}$$

For the case where $b(\theta)=\theta^{-n}$, the solution is

$$\Theta=\theta^{\frac{1}{2}(1-n)}\{CJ_{\frac{1}{2}(n-1)}(\eta\theta)+DN_{\frac{1}{2}(n-1)}(\eta\theta)\}. \tag{2.21}$$

As materials of zero or infinite magnetic permeability are unknown, the plane $\theta=0$ must be outside the region filled with this particular medium. This could occur in a V-shaped horn with metal walls lying in the planes $\theta=\theta_1$, $\theta=\theta_2$, where

$$0<\theta_1<\theta_2<2\pi.$$

In such a horn η will be determined from the boundary condition that

$$E_z=\omega/ic\cdot\varphi=0$$

at

$$\theta=\theta_1 \quad\text{and}\quad \theta=\theta_2.$$

From this it follows that for a horn the boundary conditions are

$$J_{\frac{1}{2}(n-1)}(\eta\theta_1)/N_{\frac{1}{2}(n-1)}(\eta\theta_1)=J_{\frac{1}{2}(n-1)}(\eta\theta_2)/N_{\frac{1}{2}(n-1)}(\eta\theta_2). \tag{2.22}$$

The differential equation for the radial factor can be written from (2.18)

$$0=R''+R'(1/r-a'/a)+\{\omega^2/c^2\cdot k_\theta a(r)c(r)-\eta^2/r^2\}R. \tag{2.23}$$

At some radius

$$r_0 \text{ let } a_0=a(r_0),\quad c_0=c(r_0) \tag{2.24}$$

and let

$$v=\omega/c\cdot(k_\theta a_0c_0)^{\frac{1}{2}}\cdot r=\omega/c\cdot(\mu_0\epsilon_0)^{\frac{1}{2}}r. \tag{2.25}$$

Then we find

$$0=\frac{d^2R}{dv^2}+\frac{dR}{dv}\cdot\left\{\frac{1}{v}-\frac{d}{dv}\log a\right\}+\left\{\frac{a(r)c(r)}{a_0c_0}-\frac{\eta^2}{v^2}\right\}R. \tag{2.26}$$

TABLE III. Propagation equations and field components for TM waves generated by a single component of 4-potential.

dx^1	dx^2	dx^3	α	β	γ	Propagation Equation & Field Components	Character of Waves.
dy	dr	$d\theta$	1	1	r	$\frac{\partial}{\partial z}\{\frac{1}{re}\frac{\partial \psi}{\partial z}\}+\frac{\partial}{\partial r}\{\frac{1}{re}\frac{\partial \psi}{\partial r}\}+\frac{\omega^2}{c^2}\frac{\mu}{r}\psi=0$	(a) TM wave propagated in direction dz in a cylindrical guide or co-axial line.
TM WAVES TYPE C						$H_\theta=-\omega/cr\cdot\psi$ $\quad D_z=i/r\cdot\frac{\partial \psi}{\partial r}$ $\quad D_r=-i/r\frac{\partial \psi}{\partial z}$	(b) TM waves propagated radially in a 360° horn. (c) TM waves propagated radially in a cavity with cylindrical walls.

Comparing this with the last equation given by Janke and Emde in their Chapter VIII,[4] Section 7*, which is, with x replaced by v

$$\frac{d^2y}{dv^2}+\frac{dy}{dv}\cdot\left(\frac{1}{v}-2u\right)+\left\{1+u^2-\frac{du}{dv}-\frac{u}{v}-\frac{p^2}{v^2}\right\}y=0 \quad (2.27)$$

it is clear the two Eqs. (2.26) and (2.27) are identical if

$$2u=d/dv\cdot\log a=(a'/a)\cdot c/\omega(\mu_0\epsilon_0)^{\frac{1}{2}}, \quad (2.28)$$

$$1+u^2-\frac{du}{dv}-\frac{u}{v}=\frac{a(r)c(r)}{a_0c_0}=\frac{a(v)c(v)}{a_0c_0}, \quad (2.29)$$

$$p^2=\eta^2. \quad (2.30)$$

Thus the solution becomes

$$R=a(r)^{\frac{1}{2}}\{CJ_\eta(\omega/c\cdot(\mu_0\epsilon_0)^{\frac{1}{2}}\cdot r)+DN_\eta(\omega/c\cdot(\mu_0\epsilon_0)^{\frac{1}{2}}r). \quad (2.31)$$

The function $c(v)=c(r)$

$$=\frac{a_0c_0}{a(r)}\left\{1-\frac{c^2}{2\omega^2\mu_0\epsilon_0}\left(\frac{a''}{a}-\frac{3}{2}\left(\frac{a'}{a}\right)^2+\frac{a'}{ra}\right)\right\}. \quad (2.32)$$

Thus we have, collecting results, for the case where this medium fills a conducting horn from $\theta=\theta_1$ to $\theta=\theta_2$,

$$\mu=a(r)\cdot\theta^{-n}, \quad (2.33)$$

$$\epsilon=(\mu_0\epsilon_0\theta^n/a(r))\{1-c^2/2\omega^2\mu_0\epsilon_0\cdot(a''/a-\tfrac{3}{2}(a'/a)^2+a'/ra)\}, \quad (2.34)$$

$$\varphi=\theta^{-q}\{J_q(\eta\theta)/J_q(\eta\theta_1)-N_q(\eta\theta)/N_q(\eta\theta_1)\}\times a(r)^{\frac{1}{2}}\cdot H_\eta^{(1)\text{ or }(2)}(\omega(\mu_0\epsilon_0)^{\frac{1}{2}}r/c),$$

$$E_z=\omega/ic\cdot\varphi, \quad B_r=1/r\partial\varphi/\partial\theta, \quad B_\theta=-\partial\varphi/\partial r, \quad (2.10)$$

where

$$q=\tfrac{1}{2}(n-1).$$

These solutions contain an arbitrary function $a(r)$.

[4] Janke and Emde, *Tables of Functions* (Dover Publications, New York, 1943), p. 147.

An Example of TE Waves of Type A Using k_r-Separability

In the case of k_r-separability, the differential equation for R is, from (2.16),

$$R''+(1/r-a'/a)R'-\lambda^2/r^2\cdot R=0. \quad (2.35)$$

This becomes a very simple equation, if the medium is constructed so that

$$a(r)=Cr^A, \quad (2.36)$$

$$a'/a=A/r, \quad (2.37)$$

for then R satisfies the homogeneous equation

$$R''+(1-A)R'/r-\lambda^2R/r^2=0, \quad (2.38)$$

so that the solution is $R=Pr^{m_1}+Qr^{m_2}$; with m_1 and m_2 the two roots of $m(m-1)+m(1-A)-\lambda^2=0$ or

$$m=\tfrac{1}{2}A\pm(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}. \quad (2.39)$$

Thus when $a(r)=Cr^A$ $c(r)=k_rC^{-1}r^{-A-2}$ the factor R is

$$R=r^{\frac{1}{2}A}\{P\exp[\log r(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}]+Q\exp[-\log r(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}]\}. \quad (2.40)$$

The Θ factor satisfies (2.17), or

$$0=\Theta''+\Theta'\cdot d/d\theta\cdot\log 1/b(\theta)+\{\lambda^2+\omega^2k_rf(\theta)b(\theta)/c^2\}\Theta. \quad (2.41)$$

Again, considering the case where $b(\theta)=\theta^{-n}$, the equation becomes

$$\Theta''+n\Theta'/\theta+\{\lambda^2+\omega^2k_rf(\theta)/c^2\theta^n\}\Theta=0. \quad (2.42)$$

As in the previous section, the factor is taken as

$$\Theta=\theta^{\frac{1}{2}(1-n)}J(\theta). \quad (2.43)$$

J then satisfies

$$0=J''+J'/\theta+\{\lambda^2+\omega^2k_rf(\theta)/c^2\theta_n-\tfrac{1}{4}(1-n)^2/\theta^2\}J. \quad (2.44)$$

The equation evidently becomes a Bessel's equation if

$$f(\theta)=L^2\theta^n+\theta^{n-2}$$

the solutions being

$$\Theta=\theta^{\frac{1}{2}(1-n)}H_p^{(1)\text{ or }(2)}(\lambda_0\theta), \tag{2.45}$$

where $p^2=\frac{1}{4}(n-1)^2-\omega^2k_r/c^2$, p and λ_0 are constants, $\lambda_0^2=\lambda^2+\omega^2k_rL^2/c^2$.

Summarizing the results of this example, in a medium where

$$\mu=Cr^A\theta^{-n}, \tag{2.46}$$

$$\epsilon=k_rC^{-1}r^{-A-2}(L^2\theta^n+\theta^{n-2}) \tag{2.47}$$

the function φ is

$$\varphi=r^{\frac{1}{2}A}\{P\exp[\log r(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}] +Q\exp[-\log r(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}]\} \times\theta^{\frac{1}{2}(1-n)}\{EH_p^{(1)}(\lambda_0\theta)+FH_p^{(2)}(\lambda_0\theta)\}, \tag{2.48}$$

$$p^2=\tfrac{1}{4}(n-1)^2-\omega^2k_r/c^2, \quad \lambda_0^2=\lambda^2+L^2\omega^2k_r/c^2. \tag{2.49}$$

For a more specific example, a curved wave guide, bounded by highly conducting metal sheets at two planes of constant z, and by two cylinders of radii r_0 and r_1, may be considered. The variable r may be written

$$r=e^\sigma$$

and to make $\varphi(r_0)=0$ and $\varphi(r_1)=0$ (so that E_z vanishes at its outer and inner wall), let

$$P=e^{i\alpha_0},$$
$$Q=-e^{-i\alpha_0}.$$

Then in this case

$$R=r^{\frac{1}{2}A}\{\exp[\sigma(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}+i\alpha_0] -\exp[-\sigma(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}-i\alpha_0]. \tag{2.50}$$

This can have zeros for real values of r or σ if

$$\lambda^2+\tfrac{1}{4}A^2<0.$$

Writing

$$\lambda^2=-m^2,$$
$$(\lambda^2+\tfrac{1}{4}A^2)^{\frac{1}{2}}=i(m^2-\tfrac{1}{4}A^2)^{\frac{1}{2}}$$

and this makes

$$R=r^{\frac{1}{2}A}\cdot 2i\cdot\sin\{(m^2-\tfrac{1}{4}A^2)^{\frac{1}{2}}\sigma+\alpha_0\}. \tag{2.51}$$

When $r=r_0$, $\sigma=\sigma_0$, and $r=r_1$, $\sigma=\sigma_1$, $R=0$, so

$$(m^2-\tfrac{1}{4}A^2)^{\frac{1}{2}}\sigma_0+\alpha_0=0,$$
$$(m^2-\tfrac{1}{4}A^2)^{\frac{1}{2}}\sigma_1+\alpha_0=l\pi, \quad l=\text{integer}$$
$$m^2-\tfrac{1}{4}A^2=l^2\pi^2/(\sigma_1-\sigma_0)^2.$$

Since the Hankel's functions contain exponential factors of the type $\exp(\pm i\lambda_0\theta)$, in this case they become proportional to

$$\exp\pm i\{\omega^2k_rL^2/c^2-l^2\pi^2/(\sigma_1-\sigma_0)^2-\tfrac{1}{4}A^2\}^{\frac{1}{2}}\cdot\theta.$$

This guide will pass all frequencies above f_c where

$$4\pi^2f_c^2k_rL^2/c^2>l^2\pi^2/(\sigma_1-\sigma_0)^2+\tfrac{1}{4}A^2.$$

It will be noted f_c depends upon the two constants A and L.

Section 3. TE Fields of Type C

The specific case considered will be one whose

$$\mu=Mr^{2k-2}e^{2\alpha z}, \tag{3.01}$$

$$\epsilon=\epsilon_0\mu_0/\mu. \tag{3.02}$$

For fields of type C

$$dx^1=dz, \quad dx^2=dr, \quad dx^3=d\theta,$$
$$\alpha=1, \quad \beta=1, \quad \gamma=r. \tag{3.03}$$

The non-vanishing field components are

$$E_\theta=\omega/icr\cdot\varphi, \; B_z=1/r\cdot\partial\varphi/\partial r, \; B_r=-1/r\cdot\partial\varphi/\partial z. \tag{3.04}$$

The propagation equation is, from Table II,

$$\frac{\partial}{\partial r}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial r}\right\}+\frac{\partial}{\partial z}\left\{\frac{1}{r\mu}\frac{\partial\varphi}{\partial z}\right\}+\frac{\omega^2\epsilon}{c^2r}\cdot\varphi=0. \tag{3.05}$$

Suppose that

$$\mu=\mu_1(r)\cdot\mu_2(z), \tag{3.06}$$

$$\epsilon=\epsilon_1(r)\cdot\epsilon_2(z), \tag{3.07}$$

$$\varphi=R(r)\cdot Z(z), \tag{3.08}$$

then

$$0=\frac{r\mu_1(r)}{R}\frac{d}{dr}\left\{\frac{1}{r\mu_1(r)}\frac{dR}{dr}\right\}+\frac{\mu_2(z)}{Z}\frac{d}{dz}\left\{\frac{1}{\mu_2(z)}\frac{dZ}{dz}\right\} +\frac{\omega^2}{c^2}\mu_1(r)\epsilon_1(r)\mu_2(z)\epsilon_2(z). \tag{3.09}$$

This equation is separable if either

$$\mu_1(r)\cdot\epsilon_1(r)=k_r \;\; (k_r\text{-separability}), \tag{3.10}$$

$$\mu_2(z)\cdot\epsilon_2(z)=k_z \;\; (k_z\text{-separability}). \tag{3.11}$$

For k_r-separability

$$\frac{r\mu_1(r)}{R}\frac{d}{dr}\left\{\frac{1}{r\mu_1(r)}\frac{dR}{dr}\right\}+\lambda_1^2=0, \tag{3.12}$$

$$0=\frac{\mu_2(z)}{Z}\frac{d}{dz}\left\{\frac{1}{\mu_2(z)}\frac{dZ}{dz}\right\}-\left\{\lambda_1^2-\frac{\omega^2}{c^2}k_r\mu_2(z)\epsilon_2(z)\right\}. \tag{3.13}$$

While for k_z-separability

$$\frac{r\mu_1(r)}{R}\frac{d}{dr}\left\{\frac{1}{r\mu_1(r)}\frac{dR}{dr}\right\}+\frac{\omega^2}{c^2}k_z\mu_1(r)\epsilon_1(r)=\lambda_2^2, \tag{3.14}$$

$$\frac{\mu_2(z)}{Z}\frac{d}{dz}\left\{\frac{1}{\mu_2(z)}\frac{dZ}{dz}\right\}+\lambda_2^2=0. \tag{3.15}$$

Up to this point no use has been made of Eq. (3.01) and (3.02), and the equations might be solved by accurate or approximate values of the functions $\mu_1(r)$, $\epsilon_1(r)$ in Eq. (3.14) or arbitrary values of $\mu_2(z)$, $\epsilon_2(z)$ in (3.13). If μ and ϵ have the values given in (3.01) and (3.02), $\mu\epsilon$ is a

constant, and the equations for the factors μ_1, μ_2, ϵ_1, and ϵ_2, become

$$\mu_1 = Mr^{2k-2}, \qquad \mu_2 = e^{2az}, \tag{3.16}$$

$$\epsilon_1 = \mu_0\epsilon_0/Mr^{2k-2}, \quad \epsilon_2 = e^{-2az}. \tag{3.17}$$

Both separability conditions are met, so either (3.12), (3.13) or (3.14), (3.15) can be used. Substituting these values in (3.12) and (3.13) gives

$$\frac{Mr^{2k-1}}{R}\frac{d}{dr}\left\{\frac{1}{Mr^{2k-1}}\frac{dR}{dr}\right\}+\lambda_1^2=0 \tag{3.18}$$

and

$$\frac{e^{2az}}{Z}\frac{d}{dz}\left\{e^{-2az}\frac{dZ}{dz}\right\}-\{\lambda_1^2-\omega^2k_r/c^2\}=0$$

or

$$0=Z''-2aZ'-\{\lambda_1^2-\omega^2\mu_0\epsilon_0/c^2\}Z. \tag{3.19}$$

This is immediately solved by

$$Z=\exp pz \tag{3.20}$$

where

$$p=a\pm i(\omega^2\mu_0\epsilon_0/c^2-\lambda_1^2-a^2)^{\frac{1}{2}}. \tag{3.21}$$

Writing $R=r^kS(r)$ Eq. 3.18 becomes

$$S''+1/r\cdot S'+\{(\lambda_1)^2-k^2/r^2\}S=0. \tag{3.22}$$

Thus it is clear that the radial function is

$$R=r^k\{AJ_k(\lambda_1r)+BN_k(\lambda_1r)\} \tag{3.23}$$

and

$$\varphi=r^k\{AJ_k(\lambda_1r)+BN_k(\lambda_1r)\}e^{pz}. \tag{3.24}$$

In this problem we started with the arbitrary distribution of μ and ϵ given by (3.01) and (3.02), and showed the solution could be expressed in terms of Bessel Functions and exponentials, a generalization of the solution applying when μ and ϵ are constant, which is

$$\varphi_0=r\{AJ_1(\lambda_1r)+BN_1(\lambda_1r)\}e^{pz}. \tag{3.25}$$

The non-vanishing field components are

$$E_\theta=-i\omega/c\cdot r^{k-1}\{AJ_k(\lambda_1r)+BN_k(\lambda_1r)\}e^{pz}, \tag{3.26}$$

$$B_r=-pr^{k-1}\{AJ_k(\lambda_1(r))+BN_k(\lambda_1r)\}e^{pz}, \tag{3.27}$$

$$B_z=\lambda_1r^{k-1}\{AJ_{k-1}(\lambda_1r)+BN_{k-1}(\lambda_1r)\}e^{pz}, \tag{3.28}$$

since Bessel functions obey

$$[dZ_k/d(\lambda_1r)]=Z_{k-1}-k/\lambda_1\cdot Z_kr^{-1}.$$

The solutions agree with the classic solutions for a region where μ and ϵ are constant, as is seen on taking $k=1$, $a=0$.

This choice of the function $\mu(r, z)$ leads to an infinite value of either μ or ϵ on the axis. For this reason it is necessary to specify that the medium does not extend all the way in to the axis. An interesting case is that of a co-axial line where the space between the conductors is filled with a medium specified by Eqs. (3.01) and (3.02). If the medium fills a co-axial line, and is bounded by highly conducting metal cylinders of radii r_0 and r_1, then the condition for the vanishing of the tangential electric field is

$$J_k(\lambda_1r_0)/N_k(\lambda_1r_0)=J_k(\lambda_1r_0)/N_k(\lambda_1r_0). \tag{3.29}$$

The roots in λ_1 of this equation determine the values of the propagation constant p. This constant is complex at all angular frequencies above

$$f_c=1/2\pi\cdot c\{(\lambda_1^2+a^2)/\mu_0\epsilon_0\}^{\frac{1}{2}}.$$

Below this frequency the propagation constant is a real number.

Section 4. TM Waves of Type C

Since to each solution of the field equations representing a TE wave $\varphi_3=\varphi(\mu, \epsilon)$ there is a corresponding TM wave generated by $\psi_3=\psi(\epsilon, \mu)$, it is possible to write down immediately the TM solutions corresponding to (3.01), (3.02), and (3.24), namely

$$\epsilon=Mr^{2k-2}e^{2az}, \tag{4.01}$$

$$\mu=\mu_0\epsilon_0/\epsilon, \tag{4.02}$$

$$\psi=r^k\{CJ_k(hr)+DN_k(hr)\}e^{pz}, \tag{4.03}$$

where

$$p=a\pm i\{\omega^2\mu_0\epsilon_0/c^2-h^2-a^2\}^{\frac{1}{2}}. \tag{4.04}$$

The non-vanishing field components are

$$H_\theta=-\omega/c\cdot r^{k-1}\{CJ_k(hr)+DN_k(hr)\}e^{pz}, \tag{4.05}$$

$$D_r=-ipr^{k-1}\{CJ_k(hr)+DN_k(hr)\}e^{pz}, \tag{4.06}$$

$$D_z=ih\cdot r^{k-1}\{CJ_{k-1}(hr)+DN_{k-1}(hr)\}e^{pz}. \tag{4.07}$$

If the medium described by Eqs. (4.01) and (4.02) fills the space between two co-axial conductors of radii r_0 and r_1, the boundary conditions $D_z(r_0)=D_z(r_1)=0$ give

$$J_{k-1}(hr_0)/N_{k-1}(hr_0)=J_{k-1}(hr_1)/N_{k-1}(hr_1). \tag{4.08}$$

These results obviously reduce to the classic results for a co-axial guide when $k=1$, $a=0$.

Section 5. TEM Waves

For the case of a co-axial line filled with a medium described by (4.01) and (4.02), TEM waves can be propagated if $D_z=0$. The fields for this case are described by

$$\partial/\partial z\{1/\epsilon\cdot\partial\psi/\partial z\}+\omega^2/c^2\cdot\mu\psi=0, \tag{2.02}$$

$$D_r=-i/r\cdot\partial\psi/\partial z, \quad H_\theta=-\omega/rc\cdot\psi. \tag{2.03}$$

For this particular problem

$$\frac{d}{dz}\left\{e^{-2az}\frac{d\psi}{dz}\right\}+\frac{\omega^2}{c^2}\cdot\mu_0\epsilon_0e^{-2az}\psi=0. \tag{5.01}$$

If

$$\psi=e^{pz},$$

$$p=a\pm i\{\omega^2\mu_0\epsilon_0/c^2-a^2\}^{\frac{1}{2}}. \tag{5.02}$$

When $\omega^2<a^2c^2/\mu_0\epsilon_0$, the propagation constant p is real. If ω is very small $p=\frac{1}{2}\omega^2\mu_0\epsilon_0/ac^2$ for waves propagated

in the positive z direction. The impedance ratio becomes, from (2.04)

$$Z=iac/\omega\epsilon\mp\{\mu_0\epsilon_0/\epsilon^2-a^2c^2/\omega^2\epsilon^2\}^{\frac{1}{2}}. \quad (5.03)$$

The $\pm$ sign indicates that the impedance ratio is different for right and left traveling waves, in magnitude as well as in sign. The lower sign refers to waves traveling in the positive z direction. (The time factor is $e^{+i\omega t}$, since $\partial_4=\omega/c$.)

For waves traveling in the positive direction

$$D_r=pAe^{pz}/ir, \quad H_\theta=-\omega Ae^{pz}/cr. \quad (5.04)$$

For a co-axial line, the characteristic impedance defined as

$$Z_0=V/I=\int_{r_0}^{r_1}E_r\cdot dr\Big/\int_0^{2\pi}H\cdot ds$$

is of greater importance than the impedance ratio $Z=E_r/H_\theta$. In this case

$$Z_0=\frac{1}{2\pi M}\left\{\left[\mu_0\epsilon_0-\frac{a^2}{\omega^2c^2}\right]^{\frac{1}{2}}+\frac{ia}{\omega c}\right\}\times\left\{\frac{r_1^{2-2k}-r_0^{2-2k}}{2-2k}\right\}e^{-2az}. \quad (5.05)$$

When $k=1$, $a=0$, this reduces to

$$Z_0(1,0)=1/2\pi\cdot(\mu_0/\epsilon_0)^{\frac{1}{2}}\log_e r_1/r_0 \quad (5.06)$$

in agreement with the classic impedance of a co-axial line.

Section 6. Waves in a Region Where ϵ Varies but μ is Constant

The previous examples have dealt with cases where both μ and ϵ were functions of position. In the Type C examples, the refractive index, $(\mu\epsilon)^{\frac{1}{2}}$, was constant, although μ and ϵ varied. Another type of example, of evident practical importance, is the case where μ is constant, but ϵ is some function of position. Referring to Table III, when $dx^1=dz$, $dx^2=dr$, $dx^3=d\theta$, the propagation equation was

$$\frac{\partial}{\partial z}\left\{\frac{1}{r\epsilon}\frac{\partial\psi}{\partial z}\right\}+\frac{\partial}{\partial r}\left\{\frac{1}{r\epsilon}\frac{\partial\psi}{\partial r}\right\}+\frac{\omega^2\mu}{c^2}\frac{1}{r}\psi=0. \quad (6.01)$$

In this case μ is constant, but ϵ may be taken to be the product of two factors

$$\epsilon=\epsilon_1(r)\cdot\epsilon_2(z). \quad (6.02)$$

Factorizing ψ into $\psi=RZ$, (6.01) becomes

$$\frac{r\epsilon_1}{R}\frac{d}{dr}\left\{\frac{1}{r\epsilon_1}\frac{dR}{dr}\right\}+\frac{\epsilon_2}{Z}\frac{d}{dz}\left\{\frac{1}{\epsilon_2}\frac{dZ}{dz}\right\}+\frac{\omega^2\mu\epsilon_1\epsilon_2}{c^2}=0. \quad (6.03)$$

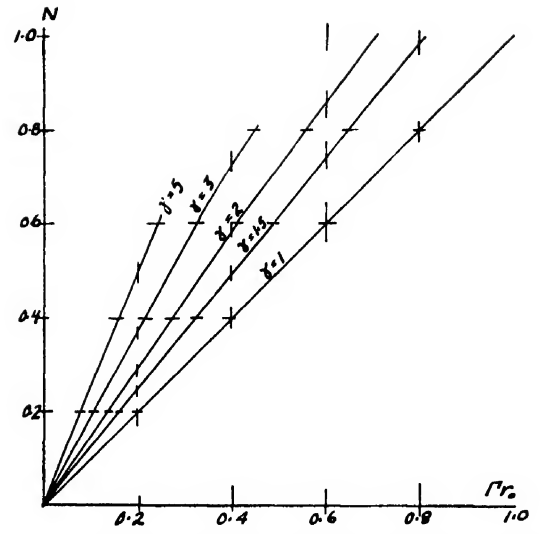


FIG. 5. Relation between the propagation constant Γ and the quantity $N=\omega r_0(\mu\epsilon_0)^{\frac{1}{2}}/c$ for a co-axial line where the ratio of outer to inner diameter is γ, filled with a medium where μ=constant, $\epsilon=\epsilon_0(r_0/r)^2$.

Applying the operator $\partial^2/\partial r\partial z$ to (6.03) gives

$$0=d\epsilon_1/dr\cdot d\epsilon_2/dz.$$

The equation is separable if ϵ is a function of r only or of z only. For co-axial line propagation, $\epsilon=\epsilon(z)$ leads to the well known results for a tapered transmission line.[5] Therefore only the case $\epsilon=\epsilon(r)$ will be considered. In this case, with $\epsilon_2=1$

$$\frac{r\epsilon(r)}{R}\frac{d}{dr}\left\{\frac{1}{r\epsilon}\frac{dR}{dr}\right\}+\left(\frac{\omega^2\mu\epsilon(r)}{c^2}+p^2\right)=0, \quad (6.04)$$

$$Z''=p^2Z. \quad (6.05)$$

The Z factor is clearly

$$Z=e^{\pm pz}. \quad (6.06)$$

The R factor depends on the chosen distribution of ϵ. For it the equation may be written

$$R''-R'\cdot d/dr\log r\epsilon+(p^2+\omega^2\mu\epsilon(r)/c^2)R=0. \quad (6.07)$$

In the particular case where $\epsilon=\epsilon_0(r_0/r)^{2}$** is chosen, the Eq. (6.07) becomes

$$R''+1/r\cdot R'+(p^2+\omega^2\mu\epsilon_0r_0^2/c^2r^2)R=0 \quad (6.08)$$

whose solution is

$$R=AJ_{i\omega r_0(\mu\epsilon_0)^{\frac{1}{2}}/c}(pr)+BN_{i\omega r_0(\mu\epsilon_0)^{\frac{1}{2}}/c}(pr). \quad (6.09)$$

This leads to a surprising exchange of role between p and $\omega^2\mu\epsilon/c^2$.

[5] Schelkunoff, *Electromagnetic Waves* (D. Van Nostrand Company, Inc. New York, 1943), p. 205.

** Again, a more general problem could be solved by taking $\omega^2\mu\epsilon/c^2=A+Br^{-2}$.

To derive an approximate expression for the relation between $\Gamma=ip$ and $N=\omega r_0/c.\ (\mu\epsilon_0)^{\frac{1}{2}}$ for low frequencies or small radii, it is necessary to use the usual positive power expansion for Bessel's functions. Since the solution contains Bessel's functions of non-integral order (order being imaginary), the solutions of the propagation equation may be rewritten as

$$\psi=\{AJ_{iN}(pr)+BJ_{-iN}(pr)\}e^{pz} \tag{6.10}$$

where

$$J_{iN}(pr)=\{\sum_0^\infty a_{2m}\cdot(pr)^{iN+2m}\}/2^{iN}\underline{|iN}, \tag{6.11}$$

$$J_{-iN}(pr)=\{\sum_0^\infty a_{2m}^*\cdot(pr)^{-iN+2m}\}/\underline{2^{-iN}|\cdot-iN} \tag{6.12}$$

and

$$a_{2m}=\underline{|Ni}/(\underline{|m|m+iN}(2i)^{2m}). \tag{6.13}$$

At the surface of a cylindrical conductor of radius r_0 the boundary condition is

$$-B/A=J_{iN}'(pr_0)/J_{-iN}'(pr_0). \tag{6.14}$$

For a co-axial line with conductors of radii r_0 and $r_1=\gamma r_0$, calling $u=pr_0$

$$J_{iN}'(u_0)/J_{-iN}'(u_0)=J_{iN}'(\gamma u_0)/J_{-iN}'(\gamma u_0). \tag{6.15}$$

Using the expansions (6.11) and (6.12) and the values of a_{2m} from (6.13) in (6.15) finally gives the boundary conditions of the co-axial line in the form

$$\sum_{s=0}^\infty u_0^{2s}A_s(N)=0 \tag{6.16}$$

with

$$A_s(N)=-i\sum_{m=0}^{s}\{a_{2m}a_{2s-2m}^*(2m+iN)\times(2s-2m-iN)\gamma^{2s-2m-iN}$$
$$-a_{2m}^*a_{2s-2m}(2m-iN)(2s-2m+iN)\gamma^{2s-2m+iN}. \tag{6.17}$$

Introducing the expression

$$\rho_{m,s}e^{i\varphi_{m,s}}=\underline{|iN|-iN}(2m+iN)\times(2s-2m-iN)/\underline{|m+iN|s-m-iN} \tag{6.18}$$

then by (6.17), (6.13), and (6.18)

$$iA_s=(-)^s2^{1-2s}\sum_{m=0}^{s}\gamma^{2s-2m}\rho_{m,s}\times\sin(\varphi_{m,s}-\sigma N)/\underline{|m|s-m} \tag{6.19}$$

with

$$\sigma=\log_e\gamma.$$

For small radii or low frequencies, we have

$$0=A_0(N)+u_0^2A_1(N) \tag{6.20}$$

leading to the approximate equation

$$-u_0^2=\Gamma^2r_0^2=4N(1+N^2)/\{(\gamma^2+1)N+(\gamma^2-1)(2+N^2)\cot\sigma N\} \tag{6.21}$$

and for extremely small values of N

$$\Gamma r_0\simeq\{2\sigma/(\gamma^2-1)\}^{\frac{1}{2}}\cdot N.$$

It is interesting to note that Γr_0 tends to the usual value for a co-axial line as $\gamma\to1$, for evidently, with $\sigma=\log\gamma$,

$$\lim\gamma\to1\ \Gamma^2r_0^2=\omega^2\mu\epsilon_0r_0^2/c^2.$$

Figure 5 shows the relation between N and Γr_0 for low frequencies.

It is hoped that these examples will serve to illustrate the application of this theory to radio and communication problems.

The author wishes to thank Hilda E. Smith for assistance with the manuscript and proof.

Hypersonic Research Facilities at the Ames Aeronautical Laboratory

VICTOR I. STEVENS
National Advisory Committee for Aeronautics, Ames Aeronautical Laboratory, Moffett Field, California
(Received June 14, 1950)

Two wind tunnels recently completed at the Ames Laboratory of the NACA— e 1 · by 14-in. supersonic wind tunnel and the supersonic free-flight wind tunnel—are capable of providing test Mach numbers well in excess of 5. Since at this time such facilities are relatively uncommon, this paper is primarily devoted to a description of these wind tunnels and their associated equipment.

The treatment of the 10- by 14-in. supersonic wind tunnel is concerned for the most part with the design and operation of the nozzle and diffuser which provides Mach numbers of 3.5 to about 8. Some mention is also made of the pressure measuring and flow visualization equipment. In the case of the supersonic free-flight wind tunnel high test Mach numbers are obtained by firing models at high speed into an oncoming supersonic air stream. Consequently the main emphasis is placed upon the launching techniques, the available range of test conditions, and the methods of obtaining data.

INTRODUCTION

ONLY a few years ago, interest in high Mach number flows was largely academic and as a result relatively little effort was directed toward solving the problems associated with generating and studying such flows. Recently, however, interest in these flows has greatly intensified and much more pressure has been brought to bear to develop research equipment which will facilitate the study of air flow at Mach numbers of 5 and above. The NACA at its Langley, Lewis, and Ames Laboratories and other research groups have been devoting increasing effort to the development of such equipment.

Perhaps the greatest obstacles to the progress of hypersonic research are the difficulties encountered in providing the high compression ratios required for the generation of high Mach number flows and in providing the instrumentation to cover the broad ranges of temperature, pressure, density, etc., that must be observed and recorded in the investigation of hypersonic flow phenomena. Two wind tunnels recently completed at the Ames Laboratory—the 10- by 14-in. supersonic wind tunnel and the supersonic free-flight wind tunnel—incorporate techniques which at least partially solve some of these problems. It is the purpose of this paper to give a brief description of these wind tunnels and their associated equipment.

10- BY 14-IN. SUPERSONIC WIND TUNNEL

The 10- by 14-in. supersonic wind tunnel has only recently been completed and is now being calibrated. It is a closed-throat, non-return, continuous-flow tunnel capable of producing any air-stream Mach number from 3.5 to something in excess of 8. The wind tunnel was designed to utilize existing compressors having a maximum exhaust pressure of only 6 atmos. absolute; consequently, it was important to keep the compression ratio required for a given Mach number as low as possible. Use of a conventional laval nozzle followed by a subsonic divergent-wall diffuser as shown in the upper half of Fig. 1 would require compression ratios of the order of 200 for a Mach number of 8. With a supply pressure of only 6 atmos. the exit pressure would have to be less than 0.03 atmos.—a value which is unacceptably low, for it would require very bulky vacuum equipment. Most of this compression ratio is required to overcome the head loss through the normal shock; therefore, use of a supersonic convergent-divergent diffuser (see lower half of Fig. 1) which decelerates the air before the normal shock, appeared advantageous to reduce this head loss and thereby reduce the compression ratio required and increase the allowable exit pressure. The limited data available at the time this design was undertaken indicated that small gains could be realized through use of this type diffuser for Mach numbers of 2 to 3, but there was little or no information available to direct the design of such a diffuser

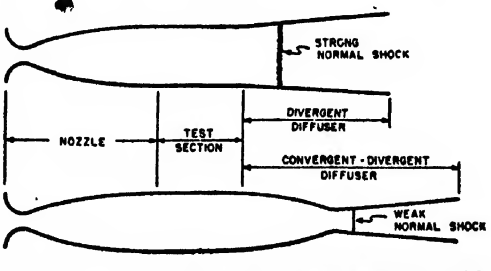


FIG. 1. Diagram of supersonic nozzles equipped with two types of diffusers.

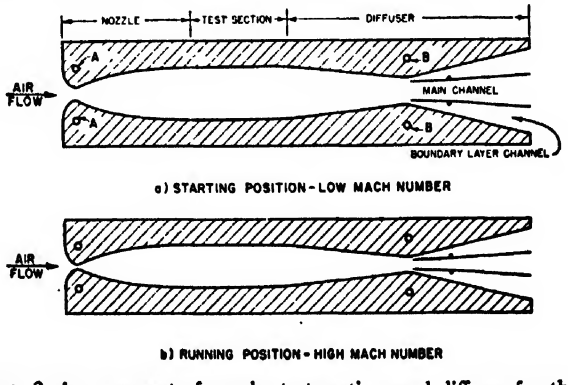


FIG. 2. Arrangement of nozzle, test section, and diffuser for the Ames 10- by 14-in. supersonic wind tunnel.

for Mach numbers as high as 8. Therefore, in 1947 an investigation of a 1- by 1-in. pilot tunnel was undertaken to develop a nozzle and diffuser for the 10- by 14-in. supersonic tunnel.

Nozzle and Diffuser

The nozzle and diffuser configuration developed is shown schematically in Fig. 2. The side walls are flat and parallel and all contour variation is in the top and bottom blocks. These blocks are rigid and are supported on pins at A and B. Jacks attached to these pins control the vertical position of the blocks and thus control the nozzle and diffuser throat openings. In a somewhat oversimplified view, opening or closing the jacks at A lowers or raises the test-section Mach number, whereas the position of jacks at B determines the location and strength of the normal shock.

Just downstream of the diffuser throat are located boundary-layer scoops. The investigations in the pilot tunnel revealed that without these scoops the normal shock boundary-layer interaction was very severe and the resulting erratic behavior of the normal shock and boundary layer made it necessary to force the normal shock far downstream of the diffuser throat in order to preserve supersonic flow ahead of the diffuser throat. The extra compression ratio required to keep the normal shock a safe distance downstream nullified, to a large extent, the gains achieved by a convergent-divergent diffuser. Removal of the boundary layer, however, stabilized the normal shock position so that supersonic flow ahead of the diffuser throat could be maintained with the shock only a short distance downstream of the scoop opening. Boundary-layer removal also greatly reduced the separation behind the normal shock and thus improved the efficiency of the subsonic diffuser.

The operation of this nozzle can probably be better understood by following through the sequence of starting supersonic flow and then changing the Mach number. The starting position is as shown in the top of Fig. 2. With the blocks open to this position, as the pressure builds up, supersonic flow at a Mach number of approximately 3.5 is established in the test section and the normal shock is pushed out of the test section through the converging section of the diffuser and into the divergent section. To establish flow with this nozzle configuration requires a compression ratio of about 6 which, as would be expected, is about the same as that required to obtain a Mach number of 3.5 in a wind tunnel with a conventional divergent diffuser. It can be shown[1] that supersonic flow in the test section will not start regardless of compression ratio available, if the second throat is not open beyond a certain minimum. For this wind tunnel the second throat is opened so that when supersonic flow is established the Mach number at the second throat is about 3. However, after flow is established the diffuser throat may be closed to give a throat Mach number of approximately 2.5. To increase the test-section Mach number, the blocks are closed as shown in the lower half of Fig. 2. With the nozzle throat closed to an opening of about 0.05 in. the test-section Mach number is 8 and the Mach number at the diffuser throat is about 3.5. The compression ratio required for this Mach number is about 30:1 which is considerably less than the estimated 200:1 required for a conventional divergent diffuser.

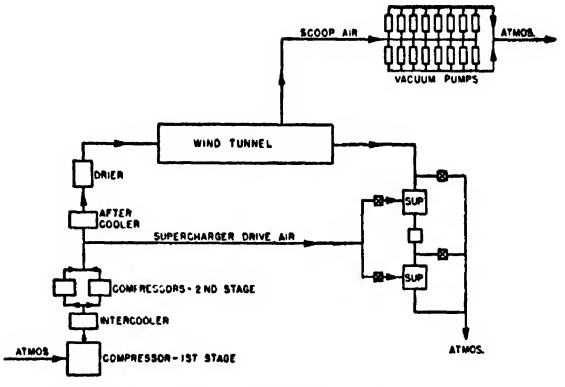


FIG. 3. Air-flow diagram for the Ames 10- by 14-in. supersonic wind tunnel.

Air Supply

The high pressure air for this wind tunnel is supplied by the equipment shown on the left side of Fig. 3. This equipment consisting of centrifugal compressors, coolers, and a silica-gel drier provides air at the nozzle entrance at a maximum total pressure of 6 atmos., a temperature of approximately 100°F and an absolute humidity of 0.0001 lb. water/lb. air. The air after passing through the nozzle and test section exhausts into either the boundary-layer scoops or the main diffuser. The low energy boundary-layer air is drawn off through the scoops by a battery of rotary vacuum pumps and discharged to the atmosphere. The air exhausting through the main diffuser is discharged directly to the atmosphere at test Mach numbers below 4.5. At Mach numbers above 4.5 this air is drawn through either one or two stages of turbine-driven superchargers to provide the required compression ratios. Since the mass flow through the tunnel decreases with increasing Mach number, it is convenient to utilize the excess air from the primary compressors to drive the turbines on the superchargers. With both superchargers operating the total compression ratio across the main passage is 55 which is more than adequate for test Mach numbers of 8.

Instrumentation

The model support and force-measuring equipment for this tunnel are very similar to that used in other supersonic wind tunnels. The models are supported

[1] A. Kantrowitz and C. du P. Donaldson, Preliminary Investigation of Supersonic Diffusers, NACA ACR L5D20 (1945).

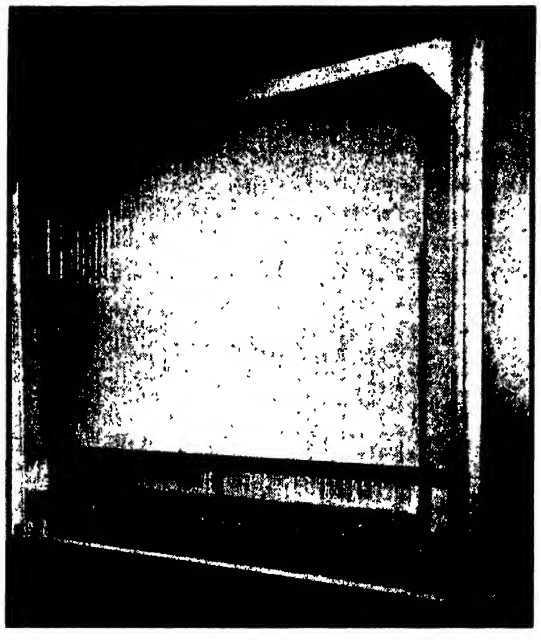


FIG. 4. Manometer board incorporating McLeod-type gauges for measuring low pressures.

from the rear on a sting which is in turn supported by a strut passing through the sidewalls. The forces and moments acting on the model are measured using a conventional strain-gauge balance system.

In contrast, the low pressures and densities existing in the wind-tunnel test section at high Mach numbers required deviations in design from the conventional pressure-measuring and flow-visualization equipment. For a given stagnation pressure the test-section static and dynamic pressures fall very rapidly with Mach number so that at a test Mach number of 8 and a stagnation pressure of 6 atmos. the static and dynamic pressures are 0.000612 and 0.0275 atmos., respectively (or 0.018 and 0.83 in. Hg). To measure pressures of this magnitude the conventional U-tube-type manometer board has been replaced by a board incorporating 40 McLeod-type gauges. This manometer board is shown in Fig. 4. Up to 40 pressures can be simultaneously recorded by photographing this board. Using these photographic records absolute pressures of 0.001 atmos. can be read to approximately one percent accuracy.

Flow about models is observed using a center-of-curvature schlieren apparatus which has much greater sensitivity to density gradients than the more common schlieren system. With the center-of-curvature schlieren the effective source and knife are located on one side of the test section at the center of curvature of the principal mirror. The mirror is located immediately adjacent to the test section on the side opposite the source and knife. Thus the light passes through the test section twice and there is no collimated light beam as in the common schlieren. The principal mirror for the 10- by 14-in. supersonic tunnel has a diameter of 1½ ft. and a radius of curvature of 50 ft. Although the schlieren system incorporating this mirror has not yet been fully checked, it is expected that the system will have adequate sensitivity over the Mach number range of this wind tunnel.

THE SUPERSONIC FREE-FLIGHT WIND TUNNEL

The supersonic free-flight wind-tunnel structure was completed early in 1949; however, the installation of the equipment required to obtain aerodynamic data was not completed until late in the fall of the same year. The major calibrations are complete and the facility is

FIG. 5. Arrangement of the Ames supersonic free-flight wind tunnel.

now being used to study the characteristics of missile-type configurations. Although the maximum test Mach number obtained to date is only slightly in excess of 8, the facility is capable of providing data for Mach numbers ranging from less than 1 to approximately 15. The necessity for high compression ratios at high test Mach numbers is obviated by generating an air-stream Mach number of only 2 or 3 in the wind-tunnel test section and providing the remainder of the total Mach number by launching the model at high velocities into the oncoming air stream. Since the air in passing through the nozzle is expanded only to a Mach number of 2 or 3, the air density and consequently the Reynolds number for a given model size are relatively high.

Data that can be obtained in this wind tunnel include the model static force characteristics which are obtained in conventional wind tunnels, as for example, lift, drag, and moment. However, the techniques for measuring these characteristics parallel those used in aeroballistic ranges.[2] The dynamic behavior of the models can also be obtained directly in this wind tunnel. The methods for obtaining these data can probably be better understood after examining the wind tunnel and related equipment.

[2] A. C. Charters, "The spark photography range," a paper given at the International Congress for Applied Mechanics (September, 1946).

Wind Tunnel and Launching Guns

The principal components of the wind tunnel proper are shown in Fig. 5. The wind tunnel is an intermittent-operation blow-down type with the air being supplied from the 12-ft. pressure wind tunnel at a maximum pressure of 6 atmos. The air flows from left to right as shown in Fig. 5 and passes, in turn, through the settling chamber, the nozzle, the test section, and the diffuser. After diffusion the air is ejected directly to the atmosphere. The test section is 1 ft. wide, 2 ft. deep, and 18 ft. long. The Mach number of the air flow in the test section is, of course, a function of the shape of the nozzle blocks. At present, the wind tunnel is equipped with two sets

(a) (b)

FIG. 6. Models and sabots as launched from guns. (a) Fin-stabilized. (b) Spin-stabilized.

of interchangeable blocks which give an air-stream Mach number of either 2 or 3. The models are fired from guns located in the diffuser and after passing through the test section and nozzle are stopped by the model catcher in the reservoir. The model catcher is a steel cylinder 2½ ft. in diameter and about 6 ft. long which is tightly packed with cotton waste backed up by wood and steel.

Launching of models at high velocity from guns requires techniques which in the past have been quite foreign to wind tunnels. The techniques which we use at present are closely patterned after those developed at the Naval Ordnance Laboratory, the Watertown Arsenal, and the Ballistic Research Laboratories at Aberdeen Proving Ground. While in the gun the models are supported in carriers or sabots illustrated in Fig. 6. The purpose of the sabot is to provide a gas seal and at the same time protect the model from the hot powder gases and to keep it properly aligned during launching. There are two basic types of sabots which we are using at present—one for fin-stabilized models

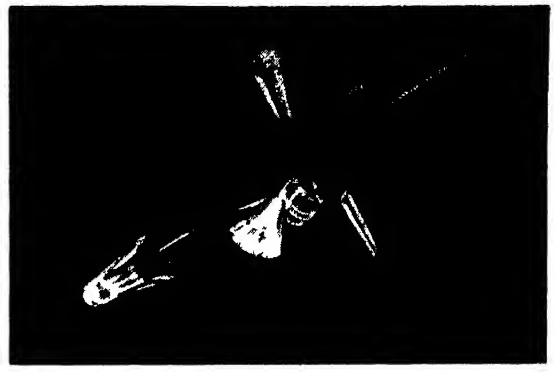


FIG. 7. Drawing showing sabot in two stages of separation from fin-stabilized model.

and one for spin-stabilized models (see Fig. 6). Both types are so constructured that when the sabot and model emerge from the gun, the sabots separate from the model and leave it free to fly on undisturbed through the wind tunnel. Sabot separation on a fin-stabilized model is illustrated in Fig. 7.

The launching guns available for use in this wind tunnel range in size from caliber .22 through .50, 20 mm, 37 mm, and 3 in. The launching velocities are varied by changing the powder charges loaded into the cartridge case—the limiting velocity being determined by the maximum pressure that the gun can withstand. In the .220 Swift, which is a sporting rifle, we have realized launching velocities of 5000 ft./sec. which, with the Mach number 2 nozzle blocks installed, gives a test Mach number of 8. It may not at first be evident how a launching velocity of 5000 ft./sec., which is a nominal sea-level Mach number of 4½, can add to the test section of 2 and yield 8. This, however, is traceable to the fact that the air in expanding through the nozzle is chilled to temperatures of about −150°F. At these low temperatures the speed of sound is of course lowered; hence, the launching velocity of 5000 ft./sec. in itself produces a Mach number of 6 which when added to the air stream Mach number of 2 gives 8. Special test barrels are being procured which should permit launching velocities of the order of 8000 ft./sec. With the Mach number 3 nozzle blocks installed in the wind tunnel this launching velocity would give a test Mach number of 15. It should be noted that a compression ratio of only 4 is required for tests at this Mach number and that to realize the same Mach number and air density in the region of the model in a conventional wind tunnel would require a stagnation pressure of more than 100,000 atmos.!

Due to the comparatively high air density in the wind-tunnel test section the test Reynolds numbers are very high for a given model size. For example, if a 6-in. model were launched in this tunnel to give a test Mach number of 7, the resulting Reynolds number would correspond to that obtained on a 50-ft. missile flying at the same Mach number at an altitude of

100,000 ft. Such a rocket is comparable in size to a V-2. In contrast, to obtain the same Reynolds number in a conventional tunnel operating at the same stagnation pressure (6 atmos.) would require a model over 10 ft. long.

Equipment and Techniques for Measuring Model Characteristics

As stated previously, the aerodynamic characteristics of the models are determined in this wind tunnel using techniques which parallel those used in aeroballistic

FIG. 8. Diagram illustrating operation of time-distance equipment.

firing ranges. To obtain drag, for example, the time-distance history of the model is recorded as the model flies through the test section and from this history the velocity, deceleration, and hence drag can be evaluated. In this wind tunnel the time-distance history is obtained using a four-station shadowgraph and a chronograph as shown in Fig. 5. As the model flies through the wind tunnel, each station in turn takes a shadowgraph of the model and a distance scale so that the distance the model travels from station to station can be measured on the shadowgraph. At the same instant each shadowgraph is taken, the chronograph records the time.

The operation of this equipment is more easily followed using Fig. 8, which isolates one of the shadowgraph stations and the principal components of the chronograph. When the model leaves the launching gun and enters the test section, it first encounters a light beam which projects through the wind tunnel and impinges on a photoelectric cell. As the model interrupts part of this light beam the output of the photoelectric cell is modified. The resulting signal is amplified and passed through a suitable time delay to allow the model to travel to the shadowgraph station. When this time delay has elapsed, the spark source is triggered and the resulting light forms a shadow image of the model on the photographic plate. Although not shown in Fig. 8, the distance scale is suspended below and along one edge of the photographic plate so that its image is also formed on the plate. When the spark is triggered, part of the light escapes from the side of the spark housing where it is picked up by a mirror and reflected along a vertical axis to a rotating prism which in turn reflects the light out to the 35-mm film on the periphery of the chronograph. Thus a latent image is formed on the chronograph film at the instant the shadowgraph is taken. As the model reaches the second, third, and fourth stations this same procedure is repeated; however, since the prism in the chronograph is rotating, the spark images are spaced approximately 90° apart on the 35-mm film. To provide a time base an H-6 mercury arc lamp is pulsed at frequencies which are regulated at any desired frequency from 10,000 to 100,000 cycles/sec. and the resulting pulsed light superimposed on the chronograph film. Since the pulse rate can be varied from 10,000 to 100,000 cycles/sec. the resulting interval between pips on the film varies from 100 to 10 μsec.

A sample of the shadowgraph and chronograph records is shown on Fig. 9. The shadowgraph was obtained on a 60° cone-cylinder model at a test Mach number of 7. The sharply defined notches of the distance scale can be seen on the upper side of the shadowgraph. At the bottom of Fig. 9 is a short segment of the chronograph record. The station marker is at the center of the section and the evenly spaced pips are the time base—each interval representing 20 μsec.

The accuracy requirements on measuring distance and time are rather stringent. For example, to restrict

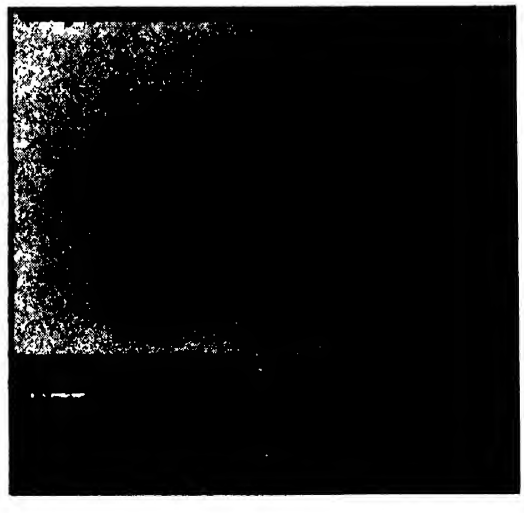


FIG. 9. Sample shadowgraph and chronograph records.

errors in drag to one percent requires in some cases distance measurements to the accuracy of 0.001 in. and time to 0.01 μsec. The requirement that time be measured to this order of accuracy becomes more significant when it is realized that in 0.01 μsec. light travels a dis-

tance of about 10 ft. It is evident, then, that in the design of equipment of this type the speed of light and light-path lengths must be considered.

Several other forms of aerodynamic data can be obtained using the time-distance record. With three lateral shadowgraph stations (omitted from Fig. 5 for clarity) and the four vertical stations already described, it is possible to define the three-dimensional flight path of the model. From this flight path, lateral and vertical accelerations, hence side force and lift, can be obtained. It is also possible to obtain base pressures using the method of characteristics to determine the conditions just ahead of the model base and using any one of the shadowgraphs to obtain the flow expansion around the base.[3] Stability parameters such as lift-curve slope, aerodynamic-center position, damping-in-roll, etc., can be evaluated by disturbing the model during the launching and observing the resulting motion. The behavior of the model during these tests can be recorded using either the shadowgraph equipment already described or by using a high speed motion-picture camera.

A sample of the data which have been obtained in this facility is given in Fig. 10 which presents the drag of a 60° cone-cylinder model as a function of Mach number. The models were launched using a .220 Swift rifle and a Lucite sabot as shown in Fig. 6. The primary purpose of the sabot for these tests was to take the rifling of the gun and thus leave the model aerodynamically smooth. As shown in Fig. 10, data were obtained for Mach numbers from 1.5 to greater than 8. Over this range the experimental data deviate from the faired curve an average of about 0.8 percent and the maximum scatter is only three percent. Since the drag of this model is relatively high, it might be expected that the scatter would be greater for better aerodynamic shapes having less drag. However, we believe that with the foreseeable improvements in equipment and measuring techniques the scatter should not be significantly increased.

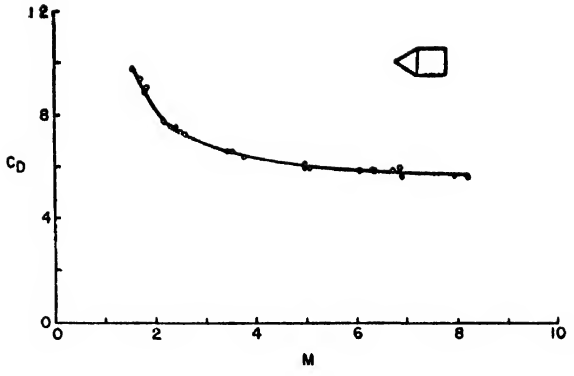


FIG. 10. Variation of drag with Mach number for 60° cone-cylinder model.

[3] A. C. Charters and R. A. Turetsky, "Determination of base pressure from free-flight data," Aberdeen Ballistic Research Laboratory (March, 1948).

The Application of Rate-Process Theory to Glass. I. Breaking Strength*

F. B. Hodgdon†
Department of Ceramics
D. A. Stuart‡ and F. E. Bjorklund
Department of Physics
University of Utah, Salt Lake City, Utah
(Received June 5, 1950)

An equation relating the ultimate strength of glass and porcelain to time under load is derived using the rate-process theory of Eyring. The data of Preston and co-workers are used to calculate the constants in the equation. The agreement between theoretical and experimental results is excellent.

I. INTRODUCTION

RECENT work of F. W. Preston and his co-workers[1-3] has provided data on the ultimate strength of glass and porcelain as a function of the time under load over a wide range of time. Because of the nature of these data it is possible to apply to ceramic materials the rate-process theory of plastic flow, developed originally by Eyring and others for textile fibers. Minor modifications of the mathematical treatment as presented by Eyring are necessary, probably because of the inherent difference between glass and fibers. These modifications do not affect the rigor of the mathematical derivation, but the physical reasons for them are still obscure and are the basis for further study.

Tobolsky and Eyring[4] have pointed out that in all unit processes of deformation, stress produces a two-fold effect on rate of strain. There is, first of all, an elastic component for which the rate of strain varies linearly with rate of stress and, secondly, a contribution to the rate of strain by relaxation processes. The latter can be readily measured for rubber, plastics, and metals, but data on ceramic materials are meager. Stress-strain, and other relationships can be derived in many ways, but it is simplest to derive them by the use of models in which the elastic component is represented by a pure spring and the relaxation process by a spring-dashpot combination.

II. PRESENT WORK

To analyze Preston's data, a model such as that illustrated in Fig. 1 was used. In this model the element shown as a spring-dashpot combination S_2-D represents a unit of the material which moves from one equilibrium position to another under the influence of an external applied force and in which bonds broken in one position are reformed in a successive position. The element shown as a pure spring, S_1 represents a network of bonds, probably primary, which once broken do not repair. (There is evidence that these bonds may repair at elevated temperatures; in this case, the pure spring should be replaced by a spring-dashpot combination.)

The spring S_1 has a modulus g_1, the spring S_2, a modulus g_2, and the dashpot constant is $\alpha=a\lambda_2/2kT$, where a is a cross-sectional area of the flow unit normal to the direction of the flow, λ_2 is the distance between equilibrium positions of the flow unit, k is the Boltzmann constant and T is the absolute temperature. A detailed discussion of this definition of α is given elsewhere.[5] In this equation $a\lambda_2$ is the volume of the hole available for flow.

When an external force f_0 is applied to the system, let the load on S_1 be f_1 and the load on S_2 and D be f_2. Let the elongation of S_1 be l_1, that of S_2, l_2 and that of D, l_3. After the load is applied and the initial elastic deformation of the system has taken place, the rate of flow of the dashpot is given by the equation

$$dl_3/dt=K\sinh\alpha f_2, \tag{1}$$

where K is the rate constant for the dashpot.[5] If the spring S_2, after the initial force has been applied, is stressed so that the force on it is f_{20}, then the force on S_2 at any time will be $f_{20}-g_2l_3$.

From this, by substitution in Eq. (1), we obtain

$$dl_3/dt=K\sinh\alpha(f_{20}-g_2l_3). \tag{2}$$

It is important to note at this point that Eq. (1) and the equation $f_2=f_{20}-g_2l_3$ are valid whether element 2 of the model is in compression or tension. Motion of the dashpot will always be in the direction of the applied force and this motion will decrease the force on the spring S_2.

If $f_{20}-g_2l_3$ is large, we can write, as a reasonably close

* This work was supported by the ONR, United States Navy Department, Research Contract No. N7 onr-451, Task Order I, Project Number NR-032-168.
Presented at the Spring meeting of the American Ceramic Society, Glass Division, April, 1948.
† Now at American Lava Company, Chattanooga, Tennessee.
‡ Now at Department of Engineering Materials, Cornell University, Ithaca, New York.
[1] T. C. Baker and F. W. Preston, J. App. Phys. 17, 162–170 (1946).
[2] T. C. Baker and F. W. Preston, J. App. Phys. 17, 170–178 (1946).
[3] J. L. Glathart and F. W. Preston, J. App. Phys. 17, 189–195 (1946).
[4] A. Tobolsky and H. Eyring, J. Chem. Phys. 11, 125–134 (1943).
[5] The derivation of this equation can be found in Glasstone, Laidler, and Eyring, *Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941). K is defined by $K=(2kT/h)\exp(-\Delta F^*/RT)$.

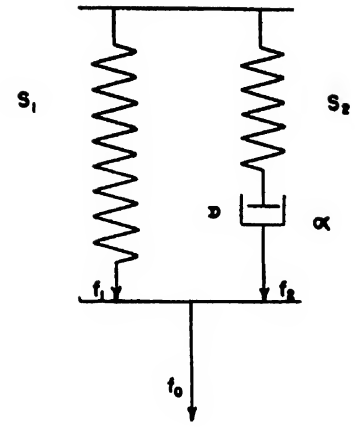


FIG. 1. Model used in derivations.

approximation

$$\frac{dl_3}{dt}=\frac{K\exp[\alpha(f_{20}-g_2l_3)]}{2}$$

which upon separation of the variables becomes

$$\frac{dl_3}{\exp[\alpha(f_{20}-g_2l_3)]}=\frac{Kdt}{2}. \tag{3}$$

Integration of Eq. (3) between the limits of $l_3=0$ when $t=0$ and $l_3=l_3$ when $t=t$ yields

$$(1/\alpha g_2)\exp[-\alpha(f_{20}-g_2l_3)]=Kt/2$$

which can be expressed in logarithmic form and arranged to give

$$f_{20}-g_2l_3=-(1/\alpha)\ln(\alpha g_2K/2)-(1/\alpha)\ln t. \tag{4}$$

Since $f_{20}-g_2l_3=f_2$ and $f_2=f_0\pm f_1$, Eq. (4) can be written

$$f_1=f_0+(1/\alpha)\ln(\alpha g_2K/2)+(1/\alpha)\ln t \text{ if } S \text{ is in tension} \tag{5a}$$

or

$$f_1=f_0-(1/\alpha)\ln(\alpha g_2K/2)-(1/\alpha)\ln t \text{ if } S \text{ is in compression.} \tag{5b}$$

This equation is equally valid with the proper sign when f_2 is a tension force and when it is a compression force. We shall consider f_2 to be a compression force and use Eq. (5b).

It has been shown by Tobolsky and Eyring[4] that with a statistical breaking of bonds, the rate of decrease of the number of primary bonds is given by the equation

$$-\frac{dN}{dt}=\frac{NkT}{h}\exp(-\Delta F^*/RT)2\sinh\frac{f_1\lambda_1}{2kTN} \tag{6}$$

where N is the number of primary bonds at time t, ΔF^* is the free energy of activation, k is the Boltzmann constant, h is Planck's constant, R is the gas constant, f_1 is the load on these bonds which causes an elongation λ_1, and T is the absolute temperature. Substitution of the value of f_1 given in Eq. (5a) in Eq. (6) gives

$$-dN/dt=(NkT/h)\times[\exp-(\Delta F_1^*/RT)]2\sinh\{[f_0-(1/\alpha)\times\ln(\alpha g_2K/2)-(1/\alpha)\ln t]\lambda_1/2kTN\}. \tag{7}$$

If the bonds do not repair after breaking, Eq. (7) may be written as an exponential:

$$-dN/dt=(NkT/h)\exp\left\{-\frac{\Delta F_1^*}{RT}+\frac{[f_0-(1/\alpha)\ln(\alpha g_2K/2)-(1/\alpha)\ln t]\lambda_1}{2kTN}\right\}. \tag{8}$$

Substituting $c=(1/\alpha)\ln(\alpha g_2K/2)$ in Eq. 8 gives

$$-dN/dt=(NkT/h)\exp(-\Delta F_1^*/RT)\times\exp[(f_0-c)\lambda_1/2kTN]t^{(-\lambda_1/2kNTa)}. \tag{9}$$

Equation (9) cannot be integrated unless further assumptions are made. In the definition $\alpha=a\lambda_2/2kT$, the numerator, $a\lambda_2$, is the average volume of the unit, the bonds of which break and repair. It is presented in the model by the dashpot. For the other bonds, those which do not repair after breaking and which are represented in the model by the pure springs, the volume is λ_1/N. As an approximation, we assume that the ratio of the volume of the non-repairable bonds to the volume of the repairable bonds is substantually constant until fracture occurs. With this assumption Eq. (9) becomes:

$$-dN/dt=(NkT/h)\exp-(\Delta F_1^*/RT)\times\exp[(f_0-e)\lambda_1/2kTN]t^{-(v_1/v_2)} \tag{10a}$$

which is equivalent to

$$\frac{dN}{(NkT/h)\exp-(\Delta F_1^*/RT)\times\exp[(f_0-c)\lambda_1/2kTN]}=-t^{-(v_1/v_2)}dt. \tag{10b}$$

When $t=0$, $N=N_0$ and when $t=t_1$, $N=0$, $v_1=\lambda_1/N_0$ and $v_2=a\lambda_2$. Equation (10b) has been integrated by Tobolsky and Eyring and gives

$$\gamma\ln t=-(f_0-c)\lambda/2kTN_0-\ln(f_0-c)\lambda/2kTN_0+(\Delta F^*-RT\ln\gamma)/RT. \tag{11}$$

III. EFFECT OF TEMPERATURE UPON BREAKING STRENGTH

If a constant load be applied to a glass specimen, it is well known that the apparent strength depends upon the temperatures. That is to say that if a constant load is applied to a specimen of glass at various temperatures the glass requires different times to fracture. Further if the time required to fracture a specimen of glass at constant load be measured and plotted *versus* the temperature, then a minimum is observed in the plotted curve.

We shall assume that v_1 and v_2 are approximately

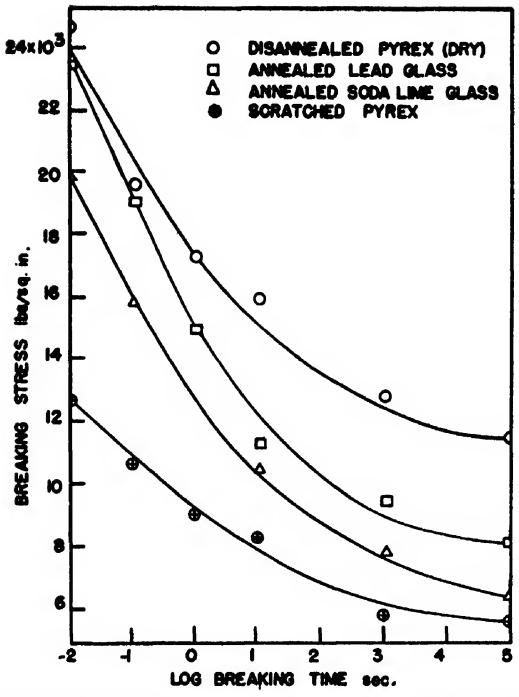


FIG. 2. Comparison of theory and experiment, breaking strength *vs.* temperature for certain glasses (after Glathart and Preston).

independent of temperature, and that $\Delta F^* = \Delta H^* - T\Delta S^*$ where ΔH^* and ΔS^* are not temperature dependant. If a minimum occurs in the time-temperature relationship, the temperature at which this minimum occurs is that temperature at which $dt/dT = 0$. Hence, differentiating Eq. (11) with respect to temperature and setting $dt/dT = 0$, we find

$$0 = \frac{\gamma}{t}\frac{dt}{dT} = \frac{(f_0 - c)v_1}{2kT}\left(\frac{1}{T} + \frac{1}{f_0 - c}\right)\frac{dc}{dt} + \frac{1}{f_0 - c}\frac{dc}{dT} - \frac{\Delta H_1^*}{RT^2} \quad (12a)$$

or

$$\frac{(f_0 - c)v_1}{2kT^2} + \left(\frac{v_1}{2kT} + \frac{1}{f_0 - c}\right)\frac{dc}{dT} - \frac{\Delta H_1^*}{RT^2} = 0 \quad (12b)$$

now

$$c = (1/\alpha)\ln(\alpha g_2 K/2)$$

or

$$c = (2kT/v_2)\ln(v_2 g_2/2h) - (2k\Delta H_2^*/Rv_2) + (2kT\Delta S_2^*/Rv_2) \quad (13)$$

and

$$dc/dT = (2k/v_2)\ln(v_2 g_2/2h) + (2k\Delta S_2^*/Rv_2). \quad (14)$$

Thus after substituting the value of dc/dT from Eq. (14) into Eq. (12a) and multiplying through by $(f_0 - c)2kT^2$ we find

$$(f_0 - c)(v_1 f_0 + (2k\Delta H_1^*\gamma/R)) + 2kT[c + (2k\Delta H_2^*/Rv_2)] = 0 \quad (15)$$

or substituting for c and rearranging

$$T^2 - \left(\frac{v_1 f_0}{2k} + \frac{2\Delta H_1^*\gamma}{R}\right)T + [(v_1 f_0/2k + \Delta H_1^*\gamma/R) \times (f_0 v_1/2k + \Delta H_2^*/R)/(\ln(v_2 g_2/2h) + \Delta S_2^*/R)] = 0.$$

Solving for T by the quadratic formula, we find,

$$T = \frac{v_1 f_0}{4k} + \frac{\Delta H_1^*\gamma}{R} \pm \left\{\left(\frac{v_1 f_0}{4k} + \frac{\Delta H_1^*\gamma}{R}\right)^2 - [(v_1 f_0/2k + \Delta H_1^*/R)(f_0 v_2/2k + \Delta H_2^*/R)/(\ln(v_2 g_2/2h) + \Delta S_2^*/R)]\right\}^{\frac{1}{2}}. \quad (16)$$

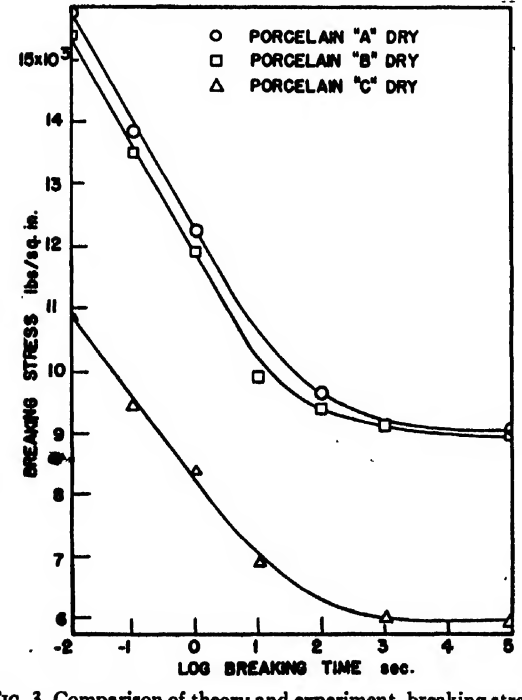


FIG. 3. Comparison of theory and experiment, breaking strength *vs.* temperature for certain porcelains (after Glathart and Preston).

It is obvious that at both the above temperatures we cannot possibly observe minima. To find at which temperature a minimum occurs, we solve for d^2t/dT^2 and substitute the above temperatures. If $d^2t/dT^2 > 0$, then at that point we observe a minimum. When this is done we find that at the lower temperature we have a minimum and at the higher a maximum.

IV. RESULTS AND DISCUSSION

The data of Preston have been used to calculate the parameters in Eq. (10a) and Figs. 2 and 3 show the results. The smooth curves are plots of the equation

TABLE I. Values of constants in the breaking strength equation.

Material	γ	ΔF_1^* kilocalories	λ_1/N_0 in^3	c psi
Disannealed Pyrex (dry)	0.272	17.1*	5.56×10^{-24}	11,000
Scratched Pyrex (wet)	0.271	17.2*	11.4×10^{-24}	5,500
Annealed lead (wet)	0.375	17.6*	7.3×10^{-24}	7,500
Annealed soda lime (wet)	0.239	16.9*	4.08×10^{-24}	5,700
Porcelain *A*	1.000	19.9	78.9×10^{-24}	9,100
Porcelain *B*	1.000	19.9	78.3×10^{-24}	8,850
Porcelain *C*	1.000	19.6	98.9×10^{-24}	8,150

* ΔF_1^* as tabulated is the apparent ΔF_1^* and should be increased by $0.5961\pi(1/\gamma)$.

when the values listed in Table I are used, and it appears that the breaking equation fits the data rather well. Further, the values obtained for the free energy of activation appear reasonable, since they are of the same order of magnitude as the strengths of alkali-oxygen and alkaline earth bond strengths in glass as determined by other methods.[6] However, as previously pointed out, a model was used in which a dashpot associated with spring S_1 has been neglected because the constant of this dashpot is such that it does affect the results at temperatures below 300°C, approximately. The simplified model presented here cannot be used in analyses of the temperature dependence of properties such as viscosity and birefringence which are now under way in these laboratories.

Data were not available with which to check reliably the breaking strength temperature results. However, calculations on a typical glass yield approximately 300°K for the minimum and 4000°K for the maximum strengths, respectively. Obviously the second temperature is so high that the maximum would never be observed.

It was assumed in our derivation that there was only an initial elongation when a load was applied, which is in accord with the known absence of "creep" in glass at ordinary temperatures; the theoretical admissability of this assumption will depend upon the slope of potential energy barrier for glass forming oxides.

The breaking equation predicts that failures in glass will be failures in tension, which is, we believe, in accord with Preston's views.

The brittle nature of glass can be explained in a qualitative manner by these results. If the strength of the alkali-oxygen and alkaline earth-oxygen bonds is of the same order of magnitude as the free energy of activation of the viscous element bond, a force sufficient to start viscous flow will also be sufficient to rupture these primary bonds. In other words, a force sufficient to cause flow, which we believe arises from the presence of "holes" in the structure is large enough to rupture bonds which do not repair. The nature of these "holes" is at present the subject of further investigation. This condition will be altered somewhat at elevated temperatures when the influence of the dashpot on spring S_1, neglected here, becomes appreciable.

ACKNOWLEDGMENT

The authors wish to express their appreciation to the staff of the School of Mines and Mineral Industries, and to H. Eyring, Dean of the Graduate School, University of Utah, for their many suggestions which were invaluable in the preparation of this paper.

[6] Kuan-Ham Sun, J. Am. Cer. Soc. **30**, 277–281 (1947).

The Application of Rate-Process Theory to Glass. II. Viscosity*

F. B. Hodgdon†
Department of Ceramics
D. A. Stuart‡
Department of Physics
University of Utah, Salt Lake City, Utah
(Received June 5, 1950)

In previous efforts to derive an equation for the temperature-viscosity relation of glass it has been assumed, tacitly or otherwise, that glass is a simple liquid. Since glass is not a simple liquid, it is necessary to find a model that will, in its reaction to applied stresses, have the same characteristics as a specimen of glass. By the use of Eyring's rate-process theory, such a model is proposed and several equations relating viscosity to time and to temperature are derived. The equations include an equation that relates the viscosity of any glass specimen to time, one which defines the rate of removal of internal strain, and an equation which relates the viscosity of a specimen an infinite time after beginning of a test to the temperature of the test. This final viscosity is the viscosity usually reported in the literature. It is found that the parameters in the equation reduced to infinite time can be calculated from the chemical composition of the glass and the viscosity of any given glass can be predicted at any temperature over a very wide range. Curves and tables are included to show the agreement of experiment and theoretical values.

I. INTRODUCTION

THE equation most generally used to define the viscosity-temperature relation of glass was derived by Sheppard[1] on the basis of the Maxwell-Boltzmann Distribution law. Over a limited temperature range, usually not exceeding 300°C, the deviations between observed and calculated viscosities are not large. Over extended temperature ranges, however, such as from 900°K to 1900°K, the agreement between theoretical and experimental values is not good. In fact, with some glasses the deviation is as much as tenfold, the reason for which is readily apparent. Implicit in Sheppard's derivation is the assumption that glasses are simple liquids. In view of the extensive literature indicating that glasses are not simple liquids, it is not surprising that an equation based upon the assumption that they are does not yield results in agreement with experiment.

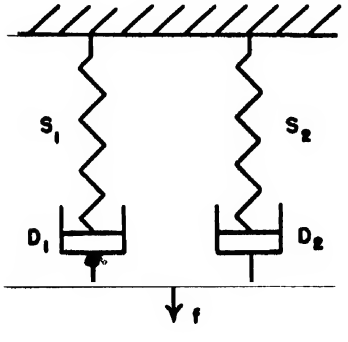


FIG. 1. Model used in derivations.

In the temperature range where the viscosity of glass is greater than 10^{10} poises, viscosities are usually measured experimentally by measuring the rate of elongation of a glass rod that is held at constant temperature and is subjected to a constant load. The quotient of the stress and the rate of elongation is, by definition, the viscosity. Data so taken can be handled readily by the type of mathematical analysis that was employed in a previous paper[2] from these laboratories to derive an equation relating time under load to the ultimate strength of glass. In that paper a model was used which consisted of pure springs and dashpots so joined that the behavior of the model under the influence of an external load represented the behavior of a glass specimen under load. It was found that a model consisting of two spring-dashpot combinations in parallel permitted the derivation of a theoretical equation for breaking strength which gave values in close agreement with experiment. The mathematical treatment of the problem was simplified by neglecting one of the dashpots; this was permissible because the characteristic of the neglected dashpot was such that it did not affect results at room temperature.

It was pointed out in that paper, however, that since the characteristic of this neglected dashpot is temperature dependent, the simplification would probably be inadmissible at higher temperatures, such as those at which the viscosity of glass is usually measured. This has been found to be the case, and in the work on viscosity, herein reported, both springs and both dashpots have been used in the mathematical derivations.

The model shown in Fig. 1 represents a glass rod subjected to a constant, external, tensile stress; since the viscosity of a given glass, as measured by a method that applies a constant tensile stress to a glass rod,

* This work was supported by the ONR, United States Navy Department, Research Contract No. N7 onr-451, Task Order 1. Project Number NR-032-168.
Presented at the Spring Meeting of the American Ceramic Society, Glass Division, April, 1949.
† Now at American Lava Company, Chattanooga, Tennessee.
‡ Now at Department of Engineering Materials, Cornell University, Ithaca, New York.
[1] S. E. Sheppard, Nature 125, 489, 709 (1930).
[2] Hodgdon, Stuart, and Bjorklund, J. App. Phys. 21, 1156 (1950).

maintained at a constant temperature, does not differ from the viscosity at that temperature, as measured by other methods, the mathematical analysis applied to this model should yield results which are applicable, regardless of the experimental method employed. It has been shown by Lillie[3] and by Dale and Stanworth[4] that the viscosity of a given glass depends, among other things, on its thermal history; a satisfactory approach to the problem of viscosity should therefore consider, first of all, the manner in which a glass specimen, maintained at constant temperature, will approach an equilibrium value for viscosity with time. This is the value for viscosity, of course, most generally reported in the literature. Therefore, the viscosity-time relation at constant temperature for a specimen initially free from internal strain, will be considered first. This is, mathematically, the simplest of the time-temperature-viscosity relations. Such freedom from strain can be attained in the laboratory by maintaining the specimen for a long period of time at a temperature very near to that at which viscosity measurements are to be made, as was done by Lillie.

The second relation to be considered will be that of the rate of removal of strain initially present in a specimen. In this case, it will not be possible to verify the rate equation which will be derived by direct comparison of predicted and experimental values, as is usually done, because there is no known way to measure, quantitatively, the amount of strain present in a sample at any time. It is true that the amount of strain may be estimated by the use of polarized light but the assumption that the amount of strain is proportional to the birefringence[5] has never been established.

The third and final relation to be analyzed is the temperature dependence of viscosity under all conditions of initial internal strain. This is the most difficult, mathematically, of the time-temperature-viscosity relations and requires the equation for the rate of strain removal for its solution. It will be possible to check the validity of the assumptions which will have to be made to derive this equation in two ways. First of all, by assuming that the initial strain is zero, it should be possible to reduce the general equation to the one derived for the special case of no initial strain and as a second check, a comparison between observed and predicted values of viscosity over a wide temperature range can be made. It will be shown that equilibrium viscosities are independent of the amount or nature of initial strain present in the specimen.

It will finally be shown that it is possible to correlate the values of the parameters in the equilibrium viscosity-time equation with the chemical composition of the glass. It will be shown that it is possible, given the chemical composition of the glass as determined by analysis, to predict the equilibrium viscosity at any temperature.

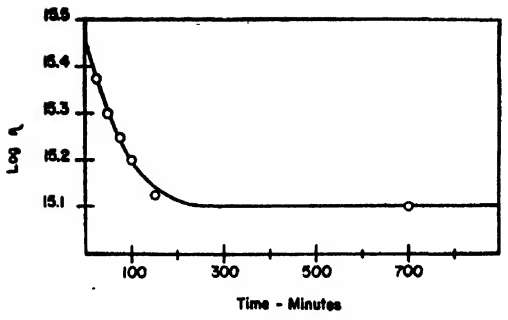


FIG. 2. Comparison of theory and experiment, log viscosity *vs.* time, for annealed glass at 486.7°C (after Lillie, see reference 3).

Two limitations are placed upon the comparison of calculated and observed values for viscosity at any temperature: the specimen must be at uniform temperature throughout its length and the viscosity determinations must be made under conditions of constant temperature. That these conditions exist is implicit in the derivations of the time-temperature-viscosity relations and represents, in any case, only those experimental conditions which would yield reproducible results.

II. NOTATION

The model used in the mathematical analyses is shown in Fig. 1. The springs represent those bonds in the glass which are stretched to their elastic limits under the influence of an external applied force and then break. They do not reform in any manner. The dashpots represent those bonds which break and then reform under the influence of the same applied force. It is not necessary that they reform in the same manner or position as that in which they originally existed; in fact, it is highly improbable that they do. No assumptions are made regarding the nature of the forces which constitute these bonds.

It is a fundamental assumption of Eyring's rate-process theory that all flow processes, of which viscous flow is one, proceed by a series of "jumps," in which discreet units of molecular dimensions move from one equilibrium position to another under the influence of an external applied force. This motion is analogous, mathematically, to the motion of a dashpot. No attempt is made to differentiate between dashpots; they differ only in that they have different temperature dependent characteristics; that is, there will be a difference in the change in their rates of flow with temperature. Both dashpots, however, obey the hyperbolic sine law; that is, their rates of flow are proportional to the hyperbolic sine of the applied force. A complete discussion of these dashpots is given elsewhere.[6]

[3] H. R. Lillie, J. Am. Cer. Soc. **16**, 619–631 (1933).
[4] A. E. Dale and J. E. Stanworth, J. Soc. Glass Tech. **29**, 414–427 (1945).
[5] G. W. Morey, *Properties of Glass* (Reinhold Publishing Corporation, New York, 1938), pp. 342ff.
[6] Glasstone, Ladler, and Eyring, *Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941).

The following symbols are used throughout this paper.

f = the total external applied force per unit area.
f_1 = the total stress on unit 1.
f_2 = the total stress on unit 2.
$f_{1,0}$ = the total stress on unit 1 at $t=0$.
$f_{2,0}$ = the total stress on unit 2 at $t=0$.
f_1' = the stress on unit 1 due to f.
f_2' = the stress on unit 2 due to f.
f'' = the total internal stress.
f_1'' = the stress on unit 1 due to f''.
f_2'' = the stress on unit 2 due to f''.
l = the total strain (elongation per unit length).
l_1 = the total elongation of spring S_1 at any time.
l_2 = the total elongation of spring S_2 at any time.
l_1' = the elongation of spring S_1 at any time caused by f_1'.
l_2' = the elongation of spring S_2 at any time caused by f_2'.
l_1'' = the elongation of spring S_1 at any time caused by f''.
l_2'' = the elongation of spring S_2 at any time caused by f''.
g_1 = the spring constant of spring S_1 so that $f_1=g_1l_1$.
g_2 = the spring constant of spring S_2 so that $f_2=g_2l_2$.
k = Boltzmann constant—1.37×10^{-16} ergs per degree.
h = Planck's constant—6.624×10^{-27} erg seconds.
R = gas constant—1.988 calories per degree.
T = temperature in degrees Kelvin.
η = viscosity in poises.
t = time in seconds.
ΔF_1^* = free energy of activation of unit 1 in calories per mole.
ΔF_2^* = free energy of activation of unit 2 in calories per mole.
ΔH_1^* = heat of activation of unit 1 in calories per mole at absolute zero.
ΔH_2^* = heat of activation of unit 2 in calories per mole at absolute zero.
ΔS_1^* = entropy of activation of unit 1 in calories per degree.
ΔS_2^* = entropy of activation of unit 2 in calories per degree.
α_1 = "constant" of dashpot 1.
α_2 = "constant" of dashpot 2.
K_1 = "characteristic" of dashpot 1.
K_2 = "characteristic" of dashpot 2.
v_1 = volume of the "hole" necessary for flow of unit 1 in cubic centimeters.
v_2 = volume of the "hole" necessary for flow of unit 2 in cubic centimeters.
λ_1 = distance in the direction of flow between equilibrium positions of unit 1 in centimeters.
λ_2 = distance in the direction of flow between equilibrium positions of unit 2 in centimeters.
a_1 = cross-sectional area of the "hole" necessary for flow of unit 1 in square centimeters.
a_2 = cross-sectional area of the "hole" necessary for flow of unit 2 in square centimeters.

The following mathematical relations exist:

$$\alpha=v/2kT,$$
$$v=a\lambda,$$

and

$$K=(2kT/h)\exp(-\Delta F^*/RT).$$

III. VISCOSITY-TIME RELATION IN THE ABSENCE OF INTERNAL STRAIN

In Fig. 1, S_1 and S_2 are pure springs, to each of which is connected, in series, a dashpot of the kind proposed by Eyring and others; that is, the dashpots are such that their rates of flow are proportional to the hyperbolic sine of the applied force:

$$dl/dt=K\sinh\alpha f.$$

At any time t the total rate of elongation will be the sum of the rate of elongation of each spring and that of its associated dashpot,

$$dl/dt=dl_1/dt+K_1\sinh\alpha_1f_1, \tag{1a}$$

$$dl/dt=dl_2/dt+K_2\sinh\alpha_2f_2. \tag{1b}$$

Since, by Hooke's law, $f_1=g_1l_1$ and $f_2=g_2l_2$, it follows that:

$$dl_1/dt=(1/g_1)(df_1/dt) \tag{2a}$$

and

$$dl_2/dt=(1/g_2)(df_2/dt). \tag{2b}$$

Since $f_1+f_2=f$, the total external applied force per unit area, which is constant:

$$df_1/dt=-df_2/dt. \tag{3}$$

Thus after elimination of f_1, l_2, and l_1, Eqs. (1a) and (1b) become:

$$dl/dt=-(1/g_1)(df_2/dt)+K_1\sinh\alpha_1(f-f_2)$$
$$=(1/g_2)(df_2/dt)+K_2\sinh\alpha_2f_2. \tag{4}$$

If $\alpha_1(f-f_2)$ and α_2f_2 are small, as a reasonably close approximation, Eq. (4) may be written:

$$dl/dt=-(1/g_1)(df_2/dt)+K_1\alpha_1(f-f_2)$$
$$=(1/g_2)(df_2/dt)+K_2\alpha_2f_2. \tag{5}$$

If

$$m=K_1\alpha_1f,$$
$$n=K_1\alpha_1+K_2\alpha_2,$$
$$p=g_1g_2/(g_1+g_2),$$

Eq. (5) becomes

$$df_2/dt=p(m-nf_2) \tag{6}$$

which has the solution

$$f_2=m/n-De^{-npt}. \tag{7}$$

In a previous work[2] it was assumed that when the external load was applied, an immediate elastic deformation occurred and that further elongation of the model was caused by movement of the dashpots. The force on the spring, before the dashpots had started to move, was called $f_{2,0}$. Using this notation we have, at $t=0$ and $f_2=f_{2,0}$

$$D=m/n-f_{2,0}=[K_1\alpha_1f/(K_1\alpha_1+K_2\alpha_2)]-f_{2,0} \tag{8}$$

by means of which the constant of integration can be evaluated. Since $dl/dt=(1/g_2)(df_2/dt)+K_2\alpha_2f_2$ [Eq. (4)], we obtain, by substitution of the value of f_2 in Eq. (7)

$$dl/dt=K_2\alpha_2m/n+D(np/g_2-K_2\alpha_2)e^{-npt}. \tag{9}$$

Substitution of the previously assigned values of m, n, and p, in Eq. (9) gives

$$dl/dt=K_1\alpha_1K_2\alpha_2f/(K_1\alpha_1+K_2\alpha_2)$$
$$+D\{[(K_1\alpha_1+K_2\alpha_2)g_1-(g_1+g_2)K_2\alpha_2]/(g_1+g_2)\}$$
$$\times\exp[-(K_1\alpha_1+K_2\alpha_2)g_1g_2t/(g_1+g_2)].$$

Since, by definition the viscosity η is $f/dl/dt$, it follows that

$$\eta=\frac{1}{K_1\alpha_1K_2\alpha_2/(K_1\alpha_1+K_2\alpha_2)+D/f[(K_1\alpha_1g_1-K_2\alpha_2g_2)/(g_1+g_2)]\exp[-(K_1\alpha_1+K_2\alpha_2)g_1g_2t/(g_1+g_2)]}. \quad (11)$$

At equilibrium, when $t=\infty$, Eq. (11) reduces to

$$\eta=(K_1\alpha_1+K_2\alpha_2)/K_1\alpha_1K_2\alpha_2. \quad (12)$$

By the use of the equations defining K, α, and v, Eq. (12) becomes

$$\eta=[(\exp(\Delta F_1^*/RT))/v_1+(\exp(\Delta F_2^*/RT))/v_2]. \quad (13)$$

In his work on viscosity, Lillie[3] held a glass at 477.8°C for 64 hours before making measurements. From his data, the following values for the parameters in Eq. (17) have been calculated

$K_1\alpha_1K_2\alpha_2/(K_1\alpha_1+K_2\alpha_2)=8\times10^{-16}$ reciprocal poises,

$(D/f)[(K_1\alpha_1g_1-K_2\alpha_2g_2)/(g_1+g_2)]=5.2\times10^{-16}$ reciprocal poises,

$[(K_1\alpha_1+K_2\alpha_2)g_1g_2/(g_1+g_2)]=1.66\times10^{-2}$ reciprocal seconds.

With these empirical values of the parameters, the viscosity-time curve which is shown in Fig. 2 was plotted. The smooth curve represents calculated values; the points represent Lillie's experimental observations.

IV. RATE OF STRAIN REMOVAL

Assume that the springs of the model shown in Fig. 1 are stressed in some way before they are connected to their respective dashpots. No assumptions are made regarding the nature of these stresses or the manner of their origin. Further, no assumption is made that the stress in each spring is the same. These stresses will, according to Hooke's law, satisfy the equations

$$f_1''=g_1l_1'' \quad (14a)$$

and

$$f_2''=g_2l_2''. \quad (14b)$$

The total rate of elongation of the system will be given by the equation:

$$dl''/dt=dl_1''/dt+K_1\sinh\alpha_1f_1''=dl_2''/dt+K_2\sinh\alpha_2f_2''. \quad (15)$$

Since $f''=f_1''+f_2''=g_1l_1''+g_2l_2''$, Eq. (15) can be written

$$df''/dt=g_1(dl_1''/dt)+g_2(dl_2''/dt)=g_1(dl''/dt)-g_1K_1\sinh\alpha_1f_1''+g_2(dl''/dt)-g_2K_2\sinh\alpha_2f_2''. \quad (16)$$

If the internal stresses are small, that is, if the glass is not badly strained, dl''/dt will be negligible; it will be true also, as a reasonably close approximation, that the hyperbolic sine is equal to its argument. With these approximations Eq. (16) becomes:

$$df''/dt=-g_1K_1\alpha_1f_1''-g_2K_2\alpha_2f_2''. \quad (17)$$

A solution of this differential equation is

$$f_1''=(\beta_1/\alpha_1)e^{-g_1K_1\alpha_1t}, \quad (18a)$$

$$f_2''=(\beta_2/\alpha_2)e^{-g_2K_2\alpha_2t} \quad (18b)$$

where β_1 and β_2 are dimensionless constants of integration. Since $f''=f_1''+f_2''$, it follows that

$$f''=(\beta_1/\alpha_1)e^{-g_1K_1\alpha_1t}+(\beta_2/\alpha_2)e^{-g_2K_2\alpha_2t}. \quad (19)$$

If the internal forces are not small, a solution of Eq. (17) is

$$f''=(g_1+g_2)l''+(2/\alpha_1)\tanh^{-1}R_1e^{-\alpha_1g_1K_1t}+(2/\alpha_2)\tanh^{-1}R_2e^{-\alpha_2g_2K_2t}+C \quad (20)$$

where R_1, R_2 and C are constants. Equation (20) will reduce to Eq. (19), when l'' is approximately zero and $e^{-\alpha_1g_1K_1t}$ and $e^{-\alpha_2g_2K_2t}$ are small.

The above solutions to Eq. (17) are not general. They are, however, valid and sufficient for the present purpose; a complete analysis of this problem and its applications to annealing will be presented in a subsequent paper.

V. VISCOSITY-TIME RELATION IN THE PRESENCE OF INTERNAL STRAIN

Assume that strain is present in the specimen when viscosity measurements are started. This condition will be realized experimentally if measurements are made on a freshly drawn fibre, as was done by Lillie[3] and by Dale and Stanworth.[4]

When internal strain is present, the total stress on each unit of the model shown in Fig. 1 is the sum of the internal stress on that unit and that part of the external load taken by the unit. That is

$$f_1=f_1'+f_1'' \quad (21a)$$

and

$$f_2=f_2'+f_2''. \quad (21b)$$

Since $f=f_1'+f_2'$, and at constant load, $df/dt=0$, it follows that

$$df_1'/dt=-df_2'/dt. \quad (22)$$

From Eqs. (1a) and (1b):

$$dl/dt=dl_1/dt+K_1\sinh\alpha_1f_1=dl_2/dt+K_2\sinh\alpha_2f_2,$$

or, as a reasonably close approximation, if the forces are small:

$$dl/dt=(1/g_1)(df_1/dt)+K_1\alpha_1f_1=(1/g_2)(df_2/dt)+K_2\alpha_2f_2.$$

Since $df_2/dt=(df''/dt)-(df_1/dt)$, it follows that

$$(df_1/dt)(1/g_1+1/g_2)+f_1(K_1\alpha_1+K_2\alpha_2)=(l/g_2)(df''/dt)+K_2\alpha_2f''+K_2\alpha_2f. \quad (23)$$

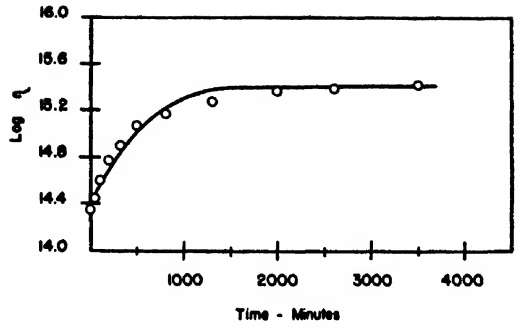


FIG. 3. Comparison of theory and experiment, log viscosity *vs.* time, for freshly drawn glass at 477.8°C (after Lillie, see reference 3).

It was shown in Eq. (19) that

$$f''=(\beta_1/\alpha_1)e^{-\alpha_1 g_1 K_1 t}+(\beta_2/\alpha_2)e^{-\alpha_2 g_2 K_2 t}.$$

Hence

$$df''/dt=-\alpha_1 g_1 K_1 f_1''-\alpha_2 g_2 K_2 f_2''$$

from which it follows that

$$df_1/dt+f_1 g_1 g_2(K_1\alpha_1+K_2\alpha_2)/(g_1+g_2)$$
$$=K_2\alpha_2 g_1 g_2 f/(g_1+g_2)+[g_1 g_2/(g_1+g_2)]$$
$$\times(K_2\alpha_2-(g_1/g_2)K_1\alpha_1)(\beta_1/\alpha_1)e^{-\alpha_1 g_1 K_1 t}. \quad (24)$$

If

$$m=(K_1\alpha_1+K_2\alpha_2)g_1 g_2/(g_1+g_2),$$
$$n=K_2\alpha_2 g_1 g_2 f/(g_1+g_2),$$
$$p=[g_1 g_2/(g_1+g_2)][K_2\alpha_2-(g_1/g_2)K_1\alpha_1](\beta_1/\alpha_1),$$

Eq. (24) can be written

$$df_1/dt+mf_1=n+pe^{-\alpha_1 g_1 K_1 t}. \quad (25)$$

For which a solution is

$$f_1=n/m+[p/(m-\alpha_1 K_1 g_1)]e^{-\alpha_1 g_1 K_1 t}+Ce^{-mt}. \quad (26)$$

When $t=0$, $f_1=f_{1,0}$ and $f_2=f_{2,0}$, where $f_{1,0}$ and $f_{2,0}$ have the meanings previously given. By means of these boundary conditions it is possible to evaluate the constant of integration in Eq. (26):

$$C=f_{1,0}-\frac{K_2\alpha_2 f}{(K_1\alpha_1+K_2\alpha_2)}-[(g_2K_2\alpha_2+g_1K_1\alpha_1)\beta_1/$$
$$(K_2\alpha_2 g_2 f-K_1\alpha_1 g_1-K_1\alpha_1 g_2)\alpha_1]. \quad (27)$$

Differentiation of Eq. (26) with respect to time gives

$$df_1/dt=-[p\alpha_1 K_1 g_1/(m-\alpha_1 K_1 g_1)]$$
$$\times e^{-\alpha_1 g_1 K_1 t}-mCe^{-mt}. \quad (28)$$

Since $dl/dt=+(1/g_1)(df_1/dt)+K_1\alpha_1 f_1$, [Eq. (5)], Eq. (28) can be written in the form:

$$dl/dt=pe^{-\alpha_1 g_1 K_1 t}$$
$$\cdot[(K_1\alpha_1-\alpha_1 g_1 K_1/g_1)/(m-\alpha_1 K_1 g_1)]$$
$$+e^{-mt}(-mC/g_1+K_1\alpha_1 C)+K_1\alpha_1 n/m. \quad (29)$$

Substitution of the values of m, n, and p, previously assigned, in Eq. (29) and simplifying gives

$$dl/dt=\frac{K_1\alpha_1 K_2\alpha_2 f}{(K_1\alpha_1+K_2\alpha_2)}+\left[\frac{C(K_1\alpha_1 g_1+K_2\alpha_2 g_2)}{(g_1+g_2)}\right]$$
$$\times\exp\left[\frac{-(K_1\alpha_1+K_2\alpha_2)g_1 g_2 t}{(g_1+g_2)}\right]. \quad (30)$$

A comparison of Eq. (30) with Eq. (10) shows that it differs only in the constant C: If the initial strain is zero, $\beta_1=0$ and

$$C=f_{1,0}-K_2\alpha_2 f/(K_1\alpha_1+K_2\alpha_2).$$

Since $f_{1,0}=f-f_{2,0}$, we can write

$$C=K_1\alpha_1 f/(K_1\alpha_1+K_2\alpha_2)-f_{2,0}$$

which is identical with the value of D in Eq. (8). At infinite time, Eq. (30) will reduce to Eq. (13).

$$\eta=h[\exp(\Delta F_1^*/RT)/v_1+\exp(\Delta F_1^*/RT)/v_2].$$

The values of the parameters in Eq. (30) for glasses having internal strain have been calculated from the data of Lillie[3] and of Dale and Stanworth.[4] With these values which are listed in Table IV, the viscosity-time relation of the glasses used have been calculated. The results are shown in Figs. 3 to 5 inclusive. The smooth curves represent theoretical values; the points are observed values.

VI. RELATION OF VISCOSITY TO COMPOSITION

An examination of the general viscosity equation [Eq. (19)] shows that it contains four temperature dependent parameters ΔF_1^*, ΔF_2^*, v_1, and v_2. In order to correlate viscosity and composition, it was necessary to establish the form of the function for each of these parameters. This cannot be done in an exact mathematical manner in the present state of the knowledge of glass structure, and it was therefore necessary to rely upon purely empirical relations. The relations which were finally adopted probably do not have any real significance; the best justification for them is that their use permits the calculation of the viscosity of a glass to a high degree of accuracy over an extended temperature range.

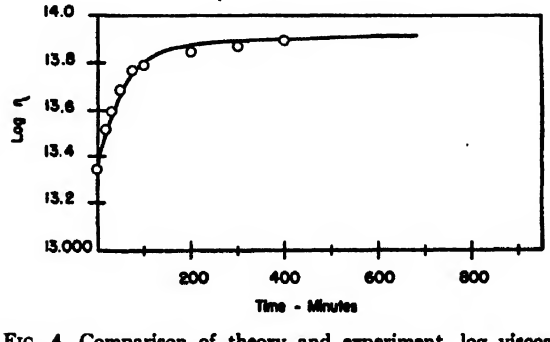


FIG. 4. Comparison of theory and experiment, log viscosity *vs.* time, for freshly drawn glass at 505.5°C (after Lillie, see reference 3).

It was first assumed that the free energies of activation were defined by the equations:

$$\Delta F_1^* = \Delta H_1^* - T\Delta S_1^*,$$
$$\Delta F_2^* = \Delta H_2^* - T\Delta S_2^*$$

and that ΔH_1^* and ΔH_2^* were independent of the temperature. In other words, it was assumed that the free energies of activation were linear functions of temperature and that the changes in them with temperature could be considered entropy changes.

A satisfactory form of the function for v_1 and v_2 was more difficult, but repeated trials finally led to the definitions:

$$v_1 = b_1 \exp(-d_1T + m_1T^2),$$
$$v_2 = b_2 \exp(-d_2T + m_2T^2)$$

where b_1, b_2 d_1, d_2, m_1, and m_2 are constants. However, it was found that unless the temperature T was very high the T^2 terms can be neglected. It was found that m_1 was of the order of magnitude of -2×10^{-8}.

Substitution of the above empirical functions in the viscosity equation, neglecting the terms in T^2, gives

$$\eta = h[\exp(\Delta H_1^*/RT - \Delta S_1^*/R + d_1T - \ln b_1) + \exp(\Delta H_2^*/RT - \Delta S_2^*/R + d_2T - \ln b_2)]$$

from which it follows, after factoring out $(\exp\Delta F_1^*/RT)/v_1$, that

$$\log\eta = (\Delta H_1^* - \Delta S_1^*T)/2.3RT + d_1T/2.3 + \log h/b_1 + \log\{1+\exp[(\Delta H_2^* - \Delta H_1^* + (\Delta S_1^* - \Delta S_2^*)T)/RT + (d_2-d_1)T + \ln b_1/b_2]\}. \quad (31)$$

If

$$M = \exp\{[\Delta H_2^* - \Delta H_1^* + (\Delta S_1^* - \Delta S_2^*)T]/RT + (d_2-d_1)T + \ln b_1/b_2\}$$

the viscosity equation can be written:

$$\log\eta = \log h + \Delta H_1^*/2.3RT - (\Delta S_1^*/2.3R + \log b_1) + d_1T/2.3 + \log(M+1) \quad (32)$$

where M has the meaning given above.

At temperatures where ΔF_1^* is much greater than ΔF_2^*, M becomes negligible and the viscosity equation reduces to

$$\log\eta = \log h + \Delta H_1^*/2.3RT - (\Delta S_1^*/2.3R + \log b_1) + d_1T/2.3. \quad (33)$$

TABLE I. Values of parameters in viscosity-time equation.

Glass designation	Temperature °C	$K_{1\alpha1}K_{1\alpha2}$ / $K_{1\alpha1}+K_{1\alpha2}$ poises^{-1}	$C(K_{1\alpha1}g_1 - K_{1\alpha2}g_2)$ / $f(g_1+g_2)$ poises^{-1}	$(K_{1\alpha1}-K_{1\alpha2})g_1g_2$ / g_1+g_2 minutes^{-1}
H. R. Lillie	447.8	1.26×10^{-14}	3.10×10^{-14}	2.41×10^{-3}
	505.5	3.80×10^{-16}	31.4×10^{-16}	3.27×10^{-3}
Dale and Stanworth	C_9 595	5.623×10^{-13}	193.9×10^{-13}	2.04×10^{-3}
	C_9 610	1.995×10^{-13}	54.2×10^{-13}	2.46×10^{-3}
	C_{40} 550	3.122×10^{-14}	398.1×10^{-14}	9.60×10^{-3}
	C_{40} 585	3.162×10^{-13}	174.6×10^{-13}	1.24×10^{-3}

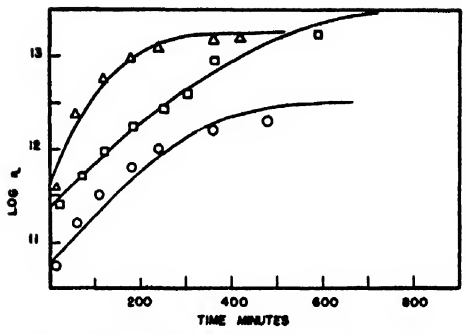


FIG. 5. Comparison of theory and experiment, log viscosity *vs.* time, for glass at different temperatures (after Dale and Stanworth, see reference 4). △ temperature 505°C, Glass No. *C* 40; □ temperature 550°C, Glass No. *C* 40; ○ temperature 595°C, Glass No. *C* 9.

This is the form of the equation which should be used when a calculation of parameters shows that ΔF_2^* is much greater than ΔF_1^*. With ordinary commercial glasses, this condition will hold when the temperature is in excess of approximately 1000°C.

An examination of the reduced viscosity equation, Eq. (33), reveals that it contains two constants, b_1 and ΔS_1^*, which are not separable. While a separation is desirable from the standpoint of an elucidation of the structure of glass, it is fortunately unnecessary if viscosities only are to be calculated. The same condition applies to the constants b_2 and ΔS_2^* in the general equation [Eq. (31)]. It is probable that this difficulty will arise, regardless of the manner in which the general viscosity equation [Eq. (19)] is expanded. For example, another expansion of that equation can be made in this manner:

$$\eta = h[(\exp(\Delta F_1^*/RT)/v_1) + (\exp(\Delta F_2^*/RT)/v_2)]$$

which can be written

$$\eta = e^{ax} + e^{by}$$

which is equivalent to

$$\eta = 2e^{(ax+by)/2}\cosh[(ax-by)/2].$$

We can then write

$$\eta = 2(h/b_1b_2)^{\frac{1}{2}}\exp\tfrac{1}{2}[(\Delta F_1^* + \Delta F_2^*)/RT + d_1T + d_2T] \times\cosh\tfrac{1}{2}[(\Delta F_1^* + \Delta F_2^*)/RT + \ln b_1/b_2 + d_1T - d_2T]$$

from which it can readily be seen that the same combinations of constants will appear.

The final assumption necessary to relate composition to viscosity is concerned with the manner in which each constituent oxide of the glass contributes to the value of each of the four parameters in the viscosity equation. It was found that the contribution of each oxide to the parameters with subscript 1 in Eqs. (31) and (33) was linearly proportional to the mole fraction of that oxide present. This is also true of the parameters with subscript 2 when alumina is not present. When it is present, it is necessary to assume that a compound with the

TABLE II. Composition of glasses used for the calculation of viscosity.

Glass designation Owen-Illinois	Composition mole percent										
	SiO_2	Al_2O_3	Fe_2O_3	B_2O_3	Na_2O	K_2O	CaO	MgO	SrO	SO_3	BaO
No. 17-1	0.7738	0.0018			0.1750				0.0478	0.0016	
17-2	0.7499	0.0006			0.1747				0.0731	0.0016	
17-3	0.7187	0.0013			0.1809				0.0975	0.0016	
17-4	0.7775	0.0123			0.1725				0.0362	0.0015	
17-5	0.7465	0.0125			0.1775				0.0620	0.0016	
17-6	0.7197	0.0133			0.1805				0.0849	0.0016	
17-7	0.7773	0.0244			0.1727				0.0240	0.0016	
17-8	0.7478	0.0249			0.1765				0.0491	0.0016	
17-9	0.7205	0.0254			0.1795				0.0230	0.0016	
15.5-1	0.7918	0.0013			0.1564				0.0490	0.0016	
15.5-2	0.7614	0.0006			0.1623				0.0749	0.0008	
15.5-3	0.7328	0.0006			0.1650				0.0999	0.0016	
15.5-4	0.7902	0.0124			0.1592				0.0367	0.0016	
15.5-5	0.7648	0.0131			0.1592				0.0621	0.0015	
15.5-6	0.7339	0.0127			0.1642				0.0876	0.0016	
15.5-7	0.7700	0.0253			0.1574				0.0372	0.0010	
15.5-8	0.7497	0.0258			0.1606				0.0629	0.0010	
15.5-9	0.7205	0.0263			0.1634				0.0887	0.0010	
14-1											
14-2	0.7793	0.0018			0.1458				0.0741	0.0010	
14-3	0.7497	0.0013			0.1474				0.1000	0.0016	
14-4	0.8092	0.0129			0.1402				0.0368	0.0008	
14-5											
14-6	0.7512	0.0127			0.1475				0.0869	0.0016	
14-7	0.7926	0.0247			0.1446				0.0372	0.0008	
14-8	0.7637	0.0258			0.1470				0.0619	0.0016	
14-9	0.7352	0.0261			0.1490				0.0872	0.0025	
10-*B*-0	0.7227	0.0006			0.1507		0.0643	0.0616			
10-*B*-1*S*	0.7149	0.0017		0.0093	0.1517		0.0622	0.0602			
10-*B*-3*S*	0.6866	0.0006		0.0265	0.1567		0.0647	0.0640			
10-*B*-5*S*	0.6808	0.0012		0.0428	0.1519		0.0627	0.0606			
10-*B*-1*D*	0.7293	0.0006		0.0086	0.1534		0.0551	0.0531			
10-*B*-3*D*	0.7367	0.0006		0.0232	0.1526		0.0438	0.0431			
10-*B*-5*D*	0.7405	0.0006		0.0425	0.1538		0.0312	0.0314			
10-*B*-1*N*	0.7248	0.0012		0.0076	0.1418		0.0631	0.0615			
10-*B*-3*N*	0.7317	0.0012		0.0230	0.1204		0.0634	0.0604			
10-*B*-5*N*	0.7305	0.0012		0.0401	0.1026		0.0636	0.0620			
12-*B*-0	0.7210	0.0012		0.0000	0.1319		0.0745	0.0715			
12-*B*-1*S*	0.7113	0.0012		0.0093	0.1331		0.0735	0.0716			
12-*B*-3*S*	0.6950	0.0012		0.0254	0.1322		0.0746	0.0716			
12-*B*-5*S*	0.6815	0.0006		0.0399	0.1305		0.0758	0.0718			
12-*B*-1*D*	0.7243	0.0012		0.0093	0.1322		0.0673	0.0658			
12-*B*-3*D*	0.7279	0.0012		0.0248	0.1343		0.0573	0.0546			
12-*B*-5*D*	0.7351	0.0012		0.0422	0.1345		0.0439	0.0431			
12-*B*-1*N*	0.7228	0.0012		0.0084	0.1221		0.0743	0.0713			
12-*B*-3*N*	0.7273	0.0012		0.0237	0.1027		0.0736	0.0716			
12-*B*-5*N*	0.7304	0.0012		0.0399	0.0830		0.0738	0.0718			
10-0	0.7227	0.0006			0.1507		0.0643	0.0616			
10-1*S*	0.7200	0.0012	0.0038		0.1518		0.0643	0.0606			
10-3*S*	0.7095	0.0012	0.0110		0.1536		0.0634	0.0613			
10-5*S*	0.6961	0.0012	0.0187		0.1563		0.0641	0.0635			
10-1*D*	0.7309	0.0012	0.0038		0.1530		0.0564	0.0548			
10-3*D*	0.7439	0.0012	0.0111		0.1557		0.0444	0.0437			
10-5*D*	0.7572	0.0012	0.0190		0.1585		0.0319	0.0322			
10-1*N*	0.7280	0.0012	0.0038		0.1429		0.0637	0.0605			
10-3*N*	0.7356	0.0012	0.0113		0.1260		0.0632	0.0626			
10-5*N*	0.7456	0.0012	0.0183		0.1089		0.0640	0.0619			

TABLE II.—*Continued.*

Glass designation Owen-Illinois	Composition mole percent										
	SiO_2	Al_2O_3	Fe_2O_3	B_2O_3	Na_2O	K_2O	CaO	MgO	SrO	SO_3	BaO
No. 12-0	0.7210	0.0012			0.1319		0.0745	0.0715			
12-1*S*	0.7163	0.0012	0.0037		0.1329		0.0740	0.0720			
12-3*S*	0.7038	0.0012	0.0109		0.1354		0.0759	0.0729			
12-5*S*	0.6915	0.0012	0.0190		0.1367		0.0766	0.0751			
12-1*D*	0.7261	0.0012	0.0037		0.1330		0.0698	0.0662			
12-3*D*	0.7389	0.0012	0.0113		0.1352		0.0581	0.0553			
12-5*D*	0.7512	0.0012	0.0192		0.1384		0.0459	0.0441			
12-1*N*	0.7163	0.0012	0.0037		0.1322		0.0741	0.0726			
12-3*N*	0.7312	0.0012	0.0112		0.1059		0.0766	0.0740			
12-5*N*	0.7409	0.0012	0.0189		0.0877		0.0764	0.0749			
Ba-0	0.7199	0.0012			0.1317		0.0744	0.0714		0.0015	0.0000
Ba-1*S*	0.7153	0.0012			0.1306		0.0759	0.0718		0.0015	0.0038
Ba-3*S*	0.7034	0.0012			0.1340		0.0756	0.0726		0.0015	0.0117
Ba-5*S*	0.6960	0.0006			0.1347		0.0755	0.0720		0.0015	0.0197
Ba-1*D*	0.7266	0.0006			0.1328		0.0686	0.0661		0.0015	0.0039
Ba-3*D*	0.7403	0.0012			0.1343		0.0570	0.0539		0.0015	0.0118
Ba-5*D*	0.7499	0.0012			0.1374		0.0459	0.0441		0.0015	0.0200
Ba-1*N*	0.7227	0.0012			0.1221		0.0749	0.0733		0.0015	0.0043
Ba-3*N*	0.7306	0.0012			0.1069		0.0756	0.0725		0.0015	0.0117
Ba-5*N*	0.7389	0.0012			0.0867		0.0775	0.0749		0.0015	0.0193
J. P. Poole											
1	0.7523				0.2477						
2	0.7601				0.1993		0.0405				
3	0.7594				0.1446		0.0960				
4	0.7568				0.0929		0.1503				
5	0.7074				0.2407		0.0519				
6	0.6629				0.2344		0.1028				
7	0.6135				0.2343		0.1522				
8	0.6979				0.2073		0.0948				
9	0.6514				0.1989		0.1496				
10	0.6126				0.1905		0.1969				
11	0.7356				0.2002			0.0462			
12	0.7724				0.1417			0.0860			
13	0.7720				0.1024			0.1255			
14	0.6969				0.2551			0.0483			
15	0.6747				0.2394			0.0859			
16	0.6035				0.2593			0.1372			
17	0.6962				0.2168			0.0870			
18	0.6481				0.2238			0.1281			
19	0.6223				0.2116			0.1661			
20	0.7455	0.0309			0.2236						
21	0.7385	0.0605			0.2010						
22	0.7214	0.0334			0.2452						
23	0.6833	0.0632			0.2535						
24	0.7352	0.0262			0.1361		0.1025				
25	0.7265	0.0559			0.1063		0.1112				
26	0.7374	0.0351			0.1507		0.0769				
27	0.7346	0.0648			0.1503		0.0503				
28	0.7122	0.0369			0.1482		0.1027				
30	0.7095	0.0377			0.2009		0.0519				
31	0.6572	0.0350			0.1996		0.1081				
32	0.6117	0.0351			0.2492		0.1040				
33	0.7572	0.0345			0.1522		0.0561				
34	0.6566	0.0364			0.1517		0.1553				
35	0.7500				0.1528		0.0521	0.0450			
36	0.6488				0.2533		0.0523	0.0458			
37	0.6640	0.0652			0.1555		0.1053				
38	0.7149	0.0207			0.1556		0.1088				
39	0.6699				0.2352				0.0950		
40	0.6346				0.2161				0.1493		
41	0.5988				0.2056				0.1956		
42	0.6771				0.2222						0.1006
43	0.6684				0.1794						0.1521
44	0.6027				0.2053						0.1920
45	0.7577				0.1233	0.1190					
46	0.7683					0.2317					

TABLE III. Comparison of observed and calculated viscosities.

Glass No. Owen-Illinois	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.
17-1	1727	2.000	1.960	1141	5.000	4.942	935	7.650	7.712
17-2	1683	2.000	1.779	1116	5.000	5.026	927	7.650	7.796
17-3	1605	2.000	2.046	1097	5.000	5.104	922	7.650	7.751
17-4	1822	2.000	1.965	1144	5.000	5.135	945	7.650	7.769
17-5	1700	2.000	2.101	1133	5.000	5.106	941	7.650	7.628
17-6	1655	2.000	2.115	1122	5.000	5.105	937	7.650	7.713
17-7	1833	2.000	2.086	1186	5.000	5.012	960	7.650	7.768
17-8	1752	2.000	2.206	1147	5.000	5.222	951	7.650	7.850
17-9	1689	2.000	2.250	1127	5.000	5.299	944	7.650	7.910
15-1	1797	2.000	1.907	1161	5.000	4.850	950	7.650	7.810
15-2	1700	2.000	1.966	1130	5.000	4.977	940	7.650	7.670
15-3	1650	2.000	1.971	1119	5.000	4.952	936	7.650	7.650
15-4	1850	2.000	2.003	1169	5.000	4.984	957	7.650	7.700
15-5	1761	2.000	2.091	1152	5.000	5.085	951	7.650	7.835
15-6	1661	2.000	2.179	1133	5.000	5.067	947	7.650	7.712
15-7	1900	2.000	2.107	1243	5.000	4.615	966	7.65	7.908
15-8	1777	2.000	2.203	1169	5.000	5.077	962	7.65	7.808
15-9	1689	2.000	2.292	1152	5.000	5.106	957	7.65	7.526
14-2	1761	2.000	1.941	1158	5.000	4.834	957	7.65	7.533
14-3	1683	2.000	2.004	1143	5.000	4.843	954	7.65	7.491
14-4	1858	2.000	2.131	1211	5.000	4.755	976	7.65	7.593
14-6	1745	2.000	2.056	1158	5.000	4.922	963	7.65	7.663
14-7	1939	2.000	2.137	1211	5.000	4.909	986	7.65	7.629
14-8	1827	2.000	2.194	1191	5.000	4.967	981	7.65	7.731
14-9	1755	2.000	2.211	1169	5.000	5.081	975	7.65	7.688
10-0	1747	2.000	2.008	1163	5.000	5.232	978	7.65	7.704
10-1*S*	1744	2.000	1.985	1155	5.000	5.232	975	7.65	7.696
10-3*S*	1700	2.000	2.074	1147	5.000	5.232	966	7.65	7.812
10-5*S*	1666	2.000	2.128	1139	5.000	5.226	961	7.65	7.804
10-1*D*	1750	2.000	1.984	1161	5.000	5.154	975	7.65	7.645
10-3*D*	1750	2.000	1.977	1161	5.000	5.042	964	7.65	7.734
10-5*D*	1783	2.000	1.881	1158	5.000	4.948	955	7.65	7.697
10-1*N*	1755	2.000	2.008	1172	5.000	5.064	983	7.65	7.666
10-3*N*	1777	2.000	2.017	1202	5.000	4.894	1000	7.65	7.495
10-5*N*	1852	2.000	1.956	1216	5.000	4.807	1014	7.65	7.400
12-0	1736	2.000	2.047	1194	5.000	5.054	1000	7.65	7.594
12-1*S*	1756	2.000	2.047	1183	5.000	5.112	994	7.65	7.678
12-3*S*	1705	2.000	2.139	1172	5.000	5.139	989	7.65	7.642
12-5*S*	1689	2.000	2.139	1164	5.000	5.142	983	7.65	7.734
12-1*D*	1755	2.000	2.047	1189	5.000	5.043	994	7.65	7.735
12-3*D*	1777	2.000	1.980	1189	5.000	4.938	986	7.65	7.606
12-5*D*	1794	2.000	1.929	1183	5.000	4.872	977	7.65	7.576
12-1*N*	1772	2.000	1.981	1191	5.000	5.049	1008	7.65	7.384
12-3*N*	1794	2.000	2.052	1222	5.000	4.869	1027	7.65	7.557
12-5*N*	1855	2.000	1.977	1244	5.000	4.629	1047	7.65	7.167
10-*B*-0	1747	2.000	2.008	1164	5.000	5.200	978	7.65	7.705
10-*B*-1*S*	1686	2.000	2.107	1152	5.000	5.252	974	7.65	7.742
10-*B*-3*S*	1636	2.000	2.072	1141	5.000	5.094	968	7.65	7.735
10-*B*-5*S*	1594	2.000	2.094	1116	5.000	5.118	963	7.65	7.970
10-*B*-1*D*	1722	2.000	2.021	1164	5.000	5.067	978	7.65	7.554
10-*B*-3*D*	1683	2.000	2.046	1144	5.000	5.055	973	7.65	7.563
10-*B*-5*D*	1666	2.000	1.919	1136	5.000	4.813	974	7.65	7.338
10-*B*-1*N*	1741	2.000	1.995	1169	5.000	5.120	981	7.65	7.751
10-*B*-3*N*	1739	2.000	1.994	1186	5.000	4.874	1002	7.65	7.585
10-*B*-5*N*	1755	2.000	1.890	1200	5.000	4.645	1014	7.65	7.567
12-*B*-0	1736	2.000	2.116	1194	5.000	5.057	1000	7.65	7.593

TABLE III.—*Continued.*

Glass No. Owen-Illinois	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.	Temp. °K	$\log\eta$ obs.	$\log\eta$ calc.
12-*B*-1*S*	1708	2.000	1.952	1177	5.000	5.989	994	7.65	7.593
12-*B*-3*S*	1666	2.000	2.092	1147	5.000	5.197	983	7.65	7.942
12-*B*-5*S*	1622	2.000	2.136	1133	5.000	5.168	979	7.65	8.057
12-*B*-1*D*	1750	2.000	2.001	1183	5.000	5.020	996	7.65	7.586
12-*B*-3*D*	1730	2.000	1.948	1161	5.000	4.999	991	7.65	7.540
12-*B*-5*D*	1716	2.000	1.839	1152	5.000	4.802	990	7.65	7.409
12-*B*-1*N*	1752	2.000	2.047	1186	5.000	5.678	1003	7.65	7.649
12-*B*-3*N*	1786	2.000	1.926	1180	5.000	4.969	1019	7.65	7.565
Ba-0	1737	2.000	2.110	1195	5.000	5.038	1001	7.65	7.535
Ba-1*S*	1738	2.000	2.077	1195	5.000	5.113	991	7.65	7.753
Ba-3*S*	1688	2.000	2.158	1168	5.000	5.180	985	7.65	7.757
Ba-5*S*	1698	2.000	2.067	1154	5.000	5.254	977	7.65	7.842
Ba-1*D*	1746	2.000	2.080	1188	5.000	5.069	994	7.65	7.580
Ba-3*D*	1753	2.000	2.000	1192	5.000	4.915	991	7.65	7.459
Ba-5*D*	1756	2.000	1.952	1183	5.000	4.895	975	7.65	7.612
Ba-1*N*	1739	2.000	2.119	1202	5.000	4.995	1008	7.65	7.491
Ba-3*N*	1770	2.000	2.064	1212	5.000	4.941	1018	7.65	7.436
Ba-5*N*	1842	2.000	1.928	1245	5.000	4.737	1050	7.65	7.009
J. P. Poole									
1	824	9.808	9.537	761	12.075	12.038			
2	850	10.077	10.107	783	12.578	12.790			
3	890	10.050	10.349	823	13.013	12.962			
4	948	10.124	9.857	851	13.798	13.494			
5	851	9.477	9.601	779	12.160	12.446			
6	845	9.820	9.936	782	12.788	12.490			
7	826	10.816	11.079	784	12.979	12.857			
8	848	10.354	11.014	784	13.357	13.040			
9	889	9.198	9.530	821	12.321	12.194			
10	889	9.417	10.069	822	12.753	12.727			
11	855	9.414	9.705	767	12.262	13.163			
16	855	10.659	11.227	767	13.910	14.797			
13	948	10.483	9.273	855	13.423	12.276			
14	805	10.183	10.177	754	12.151	12.258			
15	805	10.135	10.642	743	12.300	12.712			
16	806	10.211	10.300	753	12.441	12.339			
17	834	9.673	10.153	752	12.934	13.421			
18	854	9.462	9.414	766	13.014	12.667			
19	854	10.040	9.818	774	13.396	12.715			
20	853	9.768	9.959	768	12.879	13.320			
21	874	10.267	10.922	792	13.193	14.048			
22	853	9.565	9.458	765	12.822	12.889			
23	852	10.263	10.240	764	13.676	13.662			
24	953	9.645	9.435	854	13.218	12.934			
26	916	9.929	10.461	833	12.852	13.534			
28	915	10.052	10.830	853	12.421	13.106			
30	876	9.865	10.433	804	12.587	13.191			
31	875	10.302	10.877	827	12.278	12.716			
32	849	10.234	10.597	791	12.969	12.906			
33	915	9.788	10.277	822	12.917	13.732			
34	951	9.303	9.910	847	13.456	13.702			
35	873	10.152	10.592	800	12.672	13.925			
36	824	10.104	9.956	790	11.520	11.272			
38	916	9.719	10.017	832	12.989	13.161			
39	829	9.860	9.760	749	13.434	13.281			
40	829	10.026	10.271	766	13.060	13.116			
41	829	10.139	10.576	766	13.409	13.523			
42	795	10.148	10.329	717	13.960	14.108			
43	795	10.078	10.940	712	14.230	15.233			
44	794	9.954	10.127	714	13.952	14.047			

TABLE IV. Factors for calculation of parameters from chemical composition.

Constituent	ΔH_1	$\frac{\Delta S_1}{2.3R}+\log b_1$	$d_1/2.3$	ΔH_2	$\frac{\Delta S_2}{2.3R}+\log b_2$	$d_2/2.3$
SiO_2	77,000	−13.0	0.00395	164,000	7.26	0.00714
Al_2O_3	76,000	−30.0	0.00255	—	—	—
Fe_2O_3	800,000	207.0	0.00780	10,000	30.00	0.05400
B_2O_3	25,000	−11.0	0.00300	150,000	−17.70	−0.01500
$3Na_2O.Al_2O_3$	—	—	—	40,000	35.50	0.04200
CaO	37,000	−34.0	−0.00640	140,000	−29.10	−0.01900
MgO	37,000	−34.0	0.00640	80,000	−24.00	−0.00100
BaO	10,000	−30.0	−0.00600	136,000	1.00	−0.00400
SrO	36,000	−23.4	−0.00194	140,000	−29.20	−0.02800
Na_2O	30,000	−26.0	−0.00400	34,000	−8.72	−0.00090
SO_3	25,000	−20.0	0.00160	—	—	—

formula $3Na_2O \cdot Al_2O_3$ is formed in order to evaluate the parameters with subscript 2, and to recalculate mole fractions after making adjustment in composition which this assumption requires. The proportionality constants for the oxides are listed in Table IV. The product of this constant and the mole fraction of the corresponding oxide present will give the contribution of each oxide to each parameter.

The viscosities of a large number of glasses have been calculated by means of the above equations. The results of these calculations, together with the compositions of the glasses, expressed as mole percent, are shown in Tables II and III.

VII. DISCUSSION OF RESULTS

The agreement between observed and calculated logarithms of viscosity (Table III) is, in general, good. There are several possible reasons for the deviations which occur. An error of 0.1 percent in the chemical analysis for silica will introduce an error of almost three percent in the calculated logarithm of viscosity. It is unfortunate that this condition exists but it is probably impossible to obtain a viscosity-temperature equation which does not contain several terms, dependent upon temperature, that combine in such a way that analysis errors are greatly magnified. This condition, therefore, points to the need for extreme care in analysis if viscosities are to be calculated and indicates that analyses cannot be calculated from batch weights, as is so frequently done, if accuracy in predicting viscosities is desired.

There are other possible sources of error: it may be that the forms of the temperature dependent functions of the parameters in the viscosity equation are not correct. As has been pointed out, these forms are purely empirical and were used only because they yielded reasonable results. Further, only in the case of soda and alumina was any assumption made regarding the formation of compounds; other compounds may exist which exert an influence on the values of the parameters. It cannot be emphasized too strongly that the factors for the oxides, herein presented, are tentative values, based on the best available data on viscosity, but subject to change when better data become available or our knowledge of glass structure increases. A third possible source of error is the tacit assumption that no phase changes occur in the temperature ranges considered. If such changes do occur, it will be necessary to apply corrections to the entropy terms. It is also possible that a separation of the entropy and volume terms, ΔS and b, might allow a better calculation of viscosity values.

It should also be pointed out that by reason of the form of the model used to derive the various equations, a one-to-one ratio of units designated by the subscript 1 and the subscript 2, appears to exist. It is unlikely that this is so. The determination of the correct ratio, however, involves the derivation of the viscosity-time-temperature equations on a purely statistical basis. When work on such a derivation can be completed, it may serve to explain the discrepancies between the values of ΔF_1^*, found in this paper and the values found from breaking strength data and by other methods.

VIII. ACKNOWLEDGMENTS

Many individuals have contributed to this work. The authors wish to thank Dean H. Eyring and the staff of the Naval Research Project at the University of Utah for advice and criticism. Special appreciation is due Dr. Dan McLachlan, Department of Metallurgy, University of Utah, for his advice and encouragement and to Dr. O. G. Burch, Dr. E. F. Poncelet, Mr. Stephen Barber and others of the research staff of the Owens-Illinois Glass Company. The authors are indebted to Messrs. F. J. Bjorklund and O. L. Anderson, Department of Physics, University of Utah, for their assistance in the laborious task of calculating and checking factors for the oxides, and to Mr. James P. Poole, Pennsylvania State College, for supplying invaluable data on viscosity.

Circuit Analysis of Linear Varying-Parameter Networks

LOTFI A. ZADEH
Columbia University, New York, New York
(Received May 8, 1950)

This paper describes a theory of linear varying-parameter networks which is essentially a generalization of the familiar frequency domain theory of fixed linear networks. Such basic concepts as impedance, admittance, gain, etc., are extended to linear varying-parameter networks and their important properties are outlined. Extensions are given also of the general mesh and node equations, Thevenin's theorem, dualization, and some other relations that hold in the case of fixed networks. On the whole it is shown that many theorems, properties, and relations that hold in the case of fixed networks may be extended with proper modifications to linear varying-parameter networks.

I. INTRODUCTION

A LINEAR varying-parameter network is essentially a linear system in which one or more parameter-values are functions of time. Familiar examples of such networks are the various types of amplitude and phase modulators, variable bandwidth amplifiers, variable delay networks, etc. One may also add to this list the ordinary fixed linear networks, for these are simply a degenerate form of linear varying-parameter networks.

In mathematical terms, the behavior of a linear varying-parameter network is governed by a system of linear differential equations with time-dependent coefficients. These equations interrelate the voltages and currents in various parts of the network. A typical relation of this kind involving only two variables $u(t)$ and $v(t)$, each representing a voltage or a current, may be written in the following form:

$$[a_n(t)p^n+\cdots+a_1(t)p+a_0(t)]v(t) = [b_m(t)p^m+\cdots+b_1(t)p+b_0(t)]u(t), \quad (1)$$

where $p=d/dt$ and the a's and b's are known functions of time. The variable on the right, $u(t)$, is usually referred to as the input while the variable on the left, $v(t)$, is referred to as the output or the response to $u(t)$. The standard problem in connection with Eq. (1) is to find the response $v(t)$ to a given input $u(t)$ on the assumption that the network is unexcited prior to application of $u(t)$. The latter condition means, more precisely, that if $u(t)$ is zero for, say, $t<0$, then $v(t)$ must also be zero for $t<0$.

Most of the papers[1-7] having direct or indirect bearing on the subject of linear varying-parameter networks are concerned with various methods and techniques for obtaining approximate solutions to the problem stated above. In the present paper, however, a different aspect of the theory of linear varying-parameter networks will be considered. Specifically, the purpose is to extend to such networks the concepts, principles, procedures and, more generally, the techniques commonly used in the circuit analysis of fixed linear networks in the frequency domain. The basis for such an extension has been established in reference 6. Thus, by using the frequency analysis approach described in reference 6 it is possible to introduce into the theory of linear varying-parameter networks such familiar and useful notions as impedance, admittance, gain, phase, poles, zeros, etc.—all of these becoming time-varying quantities. It is also possible to extend to linear varying-parameter networks such well-known theorems as the superposition theorem, compensation theorem, Thevenin's theorem, etc., with practically no modifications. In addition it is possible to extend to linear varying-parameter networks the familiar and powerful techniques of dualization, impedance and frequency transformation,[8] spectrum analysis,[9] network expansion, vector representation in the complex plane, pole and zero analysis, etc.

The above list is by no means exhaustive, but is sufficient to indicate that many if not most of the concepts and techniques used in the circuit analysis of fixed networks may be extended to linear varying-parameter networks. Our objective in this paper is limited to that of extending only the more basic of such concepts and techniques. It appears that the exploration of the behavior and properties of linear varying-parameter networks from the frequency domain point of view offers a fruitful field for further study and investigations.

II. GENERAL THEORY

In order to simplify the terminology, the term "variable network" will frequently be used hereinafter in the same sense as "linear varying-parameter network."

As a preliminary to the extension of the basic frequency domain concepts such as impedance, admittance, etc., to linear varying-parameter networks, it will be necessary to introduce the notion of the so-called *system*

[1] J. R. Carson, Phys. Rev. 17, 116 (1921).
[2] K. Dahr, *Die operatorenrechnung und ihre weiterentwicklung nebst anwendungen auf systeme variabler parameter* (Pettersons Bokindustri, Stockholm, 1938).
[3] L. A. Pipes, Phil. Mag. 25, 585 (1938).
[4] C. Blanc, "Sur les equations differentielles lineaires a coefficients lentement variables," Bull. Tech. Suisse Romande 74, 185 (1948).
[5] M. C. Gray and S. A. Schelkunoff, Bell Sys. Tech. J. 27, 350 (1948).
[6] L. A. Zadeh, Proc. I.R.E. 38, 291 (1950).
[7] L. A. Zadeh, J. App. Phys. 21, 642 (1950).
[8] L. A. Zadeh, Proc. I.R.E. 38, 1339 (1950).
[9] L. A. Zadeh, Proc. I.R.E. 38, 1342 (1950).

function[10] of a variable network. Briefly, the system function of a variable network N is defined by the relation

$$H(j\omega;t)=\left.\frac{v(t)}{u(t)}\right]_{u(t)=e^{j\omega t}}. \tag{2}$$

In other words, the system function of N is a function $H(j\omega;t)$ such that $H(j\omega;t)e^{j\omega t}$ represents the response of N to $u(t)=e^{j\omega t}$. It is evident that for a fixed network $H(j\omega;t)$ reduces to $H(j\omega)$ and the above definition of $H(j\omega;t)$ reduces to the conventional definition of the system function of a fixed network.

The salient characteristic of the system function is that $H(j\omega;t)$ connects the output and input of a variable network in much the same manner as $H(j\omega)$ connects the output and input of a fixed network. Specifically, it may readily be shown that the output $v(t)$ of a variable network is given by the following relation:

$$v(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}H(j\omega;t)U(j\omega)e^{j\omega t}d\omega, \tag{3}$$

where $U(j\omega)$ is the Fourier transform of $u(t)$. It is seen that Eq. (3) is similar in form to the relation connecting the output and input of a fixed network:

$$v(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}H(j\omega)U(j\omega)e^{j\omega t}d\omega. \tag{4}$$

In view of this similarity and also of the fact that in Eq. (3) the variable t in $H(j\omega;t)$ plays the role of a parameter, it follows that the output of a variable network may be expressed in the same forms as the output of a fixed network. The more important of these forms are as follows:

(a) Fourier Transform Form

The Fourier transform equivalent of Eq. (3) is:

$$v(t)=\mathfrak{F}^{-1}\{H(j\omega;t)U(j\omega)\}, \tag{5}$$

where $\mathfrak{F}^{-1}$ represents the operation of inverse Fourier transformation, and the variable t in $H(j\omega;t)$ should be treated as if it were a constant parameter. It should be noted that while $v(t)$ is the inverse Fourier transform of $H(j\omega;t)U(j\omega)$, the latter quantity is not the Fourier transform of $v(t)$. In fact, it may readily be shown[6] that the Fourier transform of $v(t)$ is given by the relation

$$V(j\omega)=\int_{-\infty}^{\infty}\Gamma(j\omega';j\omega)U(j\omega')d\omega', \tag{6}$$

where the so-called *bi-frequency system function* $\Gamma(j\omega';j\omega)$ is the Fourier transform of $H(j\omega';t)e^{j\omega' t}$.

[10] The properties of this function are discussed in detail in reference 6.

(b) Laplace Transform Form

The Laplace transform equivalent of Eq. (3) is

$$v(t)=\mathfrak{L}^{-1}\{H(s;t)U(s)\} \tag{7}$$

where s is the complex frequency ($s=\sigma+j\omega$), $U(s)$ is the Laplace transform of $u(t)$, $\mathfrak{L}^{-1}$ represents the operation of inverse Laplace transformation, and the variable t in $H(s;t)$ should be treated as a constant parameter. For example, assume that the input is a unit impulse function (Dirac's δ-function) $\delta(t)$, i.e.,

$$u(t)=\delta(t), \tag{8}$$

and that $H(s;t)$ is of the form

$$H(s;t)=1/[s+\alpha(t)], \tag{9}$$

where $\alpha(t)$ is a given function of time. Since the Laplace transform of $\delta(t)$ is unity, it follows from Eq. (7) that

$$v(t)=\mathfrak{L}^{-1}\{1/[s+\alpha(t)]\} \tag{10}$$

and hence

$$v(t)=e^{-t\alpha(t)}, \tag{11}$$

which thus represents the response of the network to a unit impulse applied at $t=0$.

(c) Operational Form

The operational equivalent of Eq. (3) is:

$$v(t)=H(p;t)u(t), \tag{12}$$

where p is the usual differential operator. As in (a) and (b), in the process of operating by $H(p;t)$ on $u(t)$, the variable t in $H(p;t)$ should be treated as if it were a constant parameter.

From the above discussion it is seen that the system function of a variable network, particularly in its operational form, provides a very convenient means for describing the behavior of the network. Thus through use of the notion of the system function the behavior of a given variable network becomes identified with that of a linear time-dependent operator $H(p;t)$ which operates on the input to the network, $u(t)$, in accordance with the standard rules of operational calculus and thus yields the output $v(t)$. A significant advantage of the system function over other means for describing the behavior of a variable network is that once $H(j\omega;t)$ has been obtained, the response to a given input may be determined in most cases by using only a table of Laplace or Fourier transforms.

The problem of the determination of the system function of a variable network may be approached in several different ways. The most straightforward approach is to start with Eq. (1), which represents the relation between the output and input of the network. For convenience of the discussion, this relation will be rewritten in a compact form

$$L(p;t)v(t)=K(p;t)u(t), \tag{13}$$

where $L(p;t)$ and $K(p;t)$ stand respectively for the left-hand and right-hand operators in Eq. (1). Applying this relation to the case where $u(t)=e^{j\omega t}$, and making use of the fact that for this case the output is equal to $H(j\omega;t)e^{j\omega t}$, Eq. (13) gives,

$$L(p;t)\{H(j\omega;t)e^{j\omega t}\}=K(p;t)e^{j\omega t}. \tag{14}$$

By use of the identities

$$K(p;t)e^{j\omega t}\equiv e^{j\omega t}K(j\omega;t) \tag{15}$$

and

$$L(p;t)\{H(j\omega;t)e^{j\omega t}\}\equiv e^{j\omega t}L(p+j\omega;t)H(j\omega;t), \tag{16}$$

Eq. (14) reduces to the following relation:

$$L(p+j\omega;t)H(j\omega;t)=K(j\omega;t). \tag{17}$$

This relation is in effect a linear differential equation in which t is the independent variable, $j\omega$ is a fixed parameter, $K(j\omega;t)$ is the forcing function, and $H(j\omega;t)$ is the dependent variable. The problem of solving this equation for $H(j\omega;t)$ is considered in reference 6. Insofar as the purposes of the present analysis are concerned, it will be sufficient to assume that Eq. (17) can be solved for $H(j\omega;t)$ and thus yield an explicit expression for the system function of the network.

An indirect and frequently more convenient approach to the problem is provided by the techniques described in the present paper. In principle, the approach consists in expressing the system function of a given network in terms of the system functions of its constituent elements. This process is quite similar in form to the conventional circuit analysis of fixed networks and for this reason will be referred to by the same name. It should be emphasized, however, that the primary objective of this paper is not that of developing a method of obtaining an explicit expression for the system function of a variable network; rather, it is that of extending the conventional circuit analysis techniques to variable networks and, in particular, of establishing rules of combining the system functions of various basic combinations of networks.

An important rule of combination of system functions —one which will be found very useful later—concerns the manner in which successive operations by two or more system functions on a given input may be replaced by a single operation with a so-called *product* system function on the same input.

Consider two system functions $H_1(s;t)$ and $H_2(s;t)$ and let

$$v(t)=H_1(p;t)u(t) \tag{18}$$

and

$$w(t)=H_2(p;t)v(t). \tag{19}$$

These two relations may be combined into the one relation

$$w(t)=H_2(p;t)[H_1(p;t)u(t)], \tag{20}$$

which means that $w(t)$ is the result of successive operation by $H_1(p;t)$ and $H_2(p;t)$ on $u(t)$.

Now let it be assumed that Eq. (20) may be replaced by

$$w(t)=H_3(p;t)u(t) \tag{21}$$

where $H_3(p;t)$ is an as yet unknown system function such that the result of operation by $H_3(p;t)$ on $u(t)$ is identical with the result of successive operation by $H_1(p;t)$ and $H_2(p;t)$ on $u(t)$. $H_3(p;t)$ will be referred to as the *product system function* of $H_1(p;t)$ and $H_2(p;t)$.

To obtain the expression for $H_3(s;t)$ in terms of $H_1(s;t)$ and $H_2(s;t)$, let $u(t)=e^{st}$ in Eq. (20). Then from Eq. (20) the corresponding $w(t)$ is found to be

$$w(t)=H_2(p;t)[H_1(s;t)e^{st}] \tag{22}$$

or, equivalently

$$w(t)=e^{st}H_2(p+s;t)H_1(s;t). \tag{23}$$

On the other hand, substituting $u(t)=e^{st}$ into Eq. (21) we obtain

$$w(t)=H_3(s;t)e^{st}. \tag{24}$$

From comparison of Eqs. (23) and (24) it follows that the product system function of $H_1(s;t)$ and $H_2(s;t)$ is given by the operational relation

$$H_3(s;t)=H_2(p+s;t)H_1(s;t). \tag{25}$$

In this relation the variable t in $H_2(p+s;t)$ should be treated as a constant parameter and $H_1(s;t)$ should be treated as a function of time involving s as a parameter. It is evident that $H_2(p+s;t)$ should be interpreted as a *bilateral* operator, since in general $H_1(s;t)$ would not vanish for negative t.

Equation (25) may be written in a symbolic form

$$H_3=H_2*H_1 \tag{25a}$$

where the operation $*$ is distributive, associative and, in general, non-commutative. As should be expected, for the particular case of fixed networks Eq. (25) reduces to the familiar relation

$$H_3(s)=H_2(s)H_1(s). \tag{26}$$

The relation given by Eq. (25) expresses essentially the group property of system functions under multiplication. In connection with this property it is convenient to introduce at this point the notion of *inverse system functions*. Thus we shall say that two system functions $H_1(s;t)$ and $H_2(s;t)$ are the inverses of each other if and only if the associated product system function is equal to unity. The inverse of a system function $H(s;t)$ will be denoted as $H^{-1}(s;t)$ or $[H(s;t)]^{-1}$. With this notation the relation between a system function and its inverse assumes the following form:

$$H(p+s;t)H^{-1}(s;t)=H^{-1}(p+s;t)H(s;t)=1 \tag{27}$$

or, more simply,

$$H*H^{-1}=H^{-1}*H=1. \tag{27a}$$

This relation represents a generalization of the corre-

sponding relation in fixed networks, namely,

$$H(s)H^{-1}(s)=H^{-1}(s)H(s)=1. \tag{28}$$

Specific applications of the relations given above to the circuit analysis of linear varying-parameter networks will be made later. At this point it will suffice to note that the concept of the system function provides the necessary basis for the extension to variable networks of the frequency domain techniques commonly used in the analysis of fixed linear networks. In what follows we shall consider first such basic network functions as impedance, admittance, etc., and then will extend to variable networks some of the more important properties and relations that hold in the case of fixed networks.

III. BASIC NETWORK FUNCTIONS AND THEIR PROPERTIES

(a) Impedance

Consider a two-terminal variable network N. Let $v(t)$ be the voltage across the terminals of N and let $i(t)$ be the current flowing through N. The *instantaneous input impedance* of N or simply, the input impedance of N, is defined as a function $Z(j\omega;t)$ such that $Z(j\omega;t)e^{j\omega t}$ represents the output voltage $v(t)$ when the input current $i(t)$ is a complex exponential $i(t)=e^{j\omega t}$. In other words,

$$Z(j\omega;t)=\frac{v(t)}{i(t)}\Bigg]_{i(t)=e^{j\omega t}}. \tag{29}$$

It is seen that $Z(j\omega;t)$ is essentially the system function connecting $v(t)$ and $i(t)$. From this it follows immediately that the voltage $v(t)$ corresponding to a given input current $i(t)$ may be expressed in the four equivalent forms represented by Eqs. (3), (5), (7), and (12). For example, in operational form,

$$v(t)=Z(p;t)i(t), \tag{30}$$

where, as usual, the variable t in $Z(p;t)$ should be treated as if it were a constant parameter.

From the general relation represented by Eq. (17) it follows that if $v(t)$ and $i(t)$ are connected to each other by a differential equation such as Eq. (13), then $Z(s;t)$ satisfies the following differential equation

$$L(p+s;t)Z(s;t)=K(s;t). \tag{31}$$

A useful interpretation of $Z(j\omega;t)$ may be obtained by using the usual vector representation of voltages, currents, impedances, etc., in the complex plane. In such a plane $Z(j\omega;t)$ may be represented as a time-varying vector whose magnitude is $|Z(j\omega;t)|$ and whose argument is the phase of $Z(j\omega;t)$. This vector may be interpreted as the instantaneous amplitude of the output voltage $v(t)$ resulting from an input current of the form $i(t)=e^{j\omega t}$.

A very basic and quite obvious property of $Z(s;t)$ is related to the manner in which any number of impedances in series may be combined into a single impedance. Thus consider n variable networks N_1, N_2, $\cdots$, N_n in series with each other, and let the input impedances of these networks be respectively $Z_1(s;t)$, $Z_2(s;t)$, $\cdots$, $Z_n(s;t)$. Because of the linear character of system functions it follows immediately that the input impedance of the over-all network is given by:

$$Z(s;t)=Z_1(s;t)+Z_2(s;t)+\cdots+Z_n(s;t). \tag{32}$$

As a simple illustration of this relation consider a network consisting of a series connection of a variable resistance $R=R(t)$, a variable inductance $L=L(t)$, and a variable capacitance $C=C(t)$. The impedances of these elements are respectively $R(t)$, $L(t)s+\dot{L}(t)$, and $1/C(t)s$,[11] where the dot represents differentiation with respect to time. From Eq. (32) it follows that the input impedance in question is:

$$Z(s;t)=R(t)+L(t)s+\dot{L}(t)+1/[C(t)s]. \tag{33}$$

The input impedance of a variable network has many other properties in addition to those discussed above. Only a few of these will be considered in the following sections. The consideration of other properties is outside the scope of the present paper.

(b) Admittance

The notions of admittance and impedance of a two-terminal variable network are duals of each other. Thus, the input admittance of a network N is defined as a function $Y(j\omega;t)$ such that $Y(j\omega;t)e^{j\omega t}$ represents the current flowing through N when a voltage of the form $v(t)=e^{j\omega t}$ is applied across the input terminals of N. In other words,

$$Y(j\omega;t)=\frac{i(t)}{v(t)}\Bigg]_{v(t)=e^{j\omega t}}. \tag{34}$$

It is seen that $Y(j\omega;t)$ is the system function connecting $i(t)$ and $v(t)$. The dual of Eq. (30) is

$$i(t)=Y(p;t)v(t). \tag{35}$$

Also, the dual of Eq. (31) is

$$K(p+s;t)Y(s;t)=L(s;t). \tag{36}$$

Corresponding to the property of series addition of impedances there is the property of shunt addition of admittances. Thus, let $Y(s;t)$ be the admittance of a network consisting of a parallel connection of N_1, N_2, $\cdots$, N_n, the admittances of these networks being $Y_1(s;t)$, $Y_2(s;t)$, $\cdots$, $Y_n(s;t)$, respectively. It is evident that $Y(s;t)$ is given by

$$Y(s;t)=Y_1(s;t)+Y_2(s;t)+\cdots+Y_n(s;t) \tag{37}$$

which is the dual of Eq. (32).

An important mutual property of the input impedance

[11] These expressions follow from the relations $i=d(cv)/dt$ and $v=d(Li)/dt$.

and admittance of a variable network is that $Z(s;t)$ and $Y(s;t)$ are inverse system functions in the sense defined in section II. This fact follows immediately from the relations

$$v(t)=Z(p;t)i(t) \tag{30}$$

and

$$i(t)=Y(p;t)v(t). \tag{35}$$

Application of Eq. (27) to $Z(s;t)$ and $Y(s;t)$ yields the following relation between the impedance and admittance of a variable network

$$Z(p+s;t)Y(s;t)=Y(p+s;t)Z(s;t)=1 \tag{38}$$

or

$$Z*Y=Y*Z=1. \tag{39}$$

This relation represents a generalization of the familiar relation

$$Z(s)Y(s)=Y(s)Z(s)=1. \tag{39a}$$

Equation (38) may be used for determining the admittance of a network from the knowledge of its impedance and *vice-versa*. For example, consider a variable capacitor the capacitance of which is a function of time, $C=C(t)$. The impedance of such a capacitor is

$$Z(s;t)=1/[C(t)s]. \tag{40}$$

From Eq. (38) it follows that the admittance of this capacitor satisfies the operational relation

$$\frac{1}{C(t)(p+s)}\{Y(s;t)\}=1. \tag{41}$$

Multiplying both sides of Eq. (41) first by $C(t)$ and then by $(p+s)$ we obtain

$$Y(s;t)=(p+s)C(t) \tag{42}$$

or

$$Y(s;t)=C(t)s+C(t), \tag{43}$$

which thus represents the admittance of a variable capacitor.

In general, the determination of the impedance of a variable network from knowledge of its admittance, or *vice-versa*, is not an easy problem. For the purposes of the present analysis it will suffice to note that in most cases the solution of this problem reduces to the solution of an integral equation which results from the application of Eq. (3) to Eq. (38). In many practical cases it is not difficult to obtain an adequate approximation to $Z(s;t)$ when $Y(s,t)$ is given, or *vice-versa*, either directly from Eq. (38) or by using an approximate method of solution of the integral equation referred to above.

(c) Gain (Transfer Function)

Let $e_1(t)$ and $e_2(t)$ denote respectively the input and output of a four-terminal variable network N. By analogy with the definitions of impedance and admittance, the *gain* or *transfer function* of N is defined as a function $G(j\omega;t)$ such that $G(j\omega;t)e^{j\omega t}$ represents the output voltage $e_2(t)$ when the input voltage $e_1(t)$ is of the form $e_1(t)=e^{j\omega t}$. In other words,

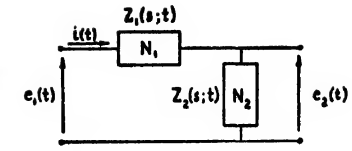


FIG. 1. Calculation of the gain of a variable network.

$$G(j\omega;t)=\left.\frac{e_2(t)}{e_1(t)}\right]_{e_1(t)=e^{j\omega t}}. \tag{44}$$

It is seen that $G(j\omega;t)$ is the system function connecting $e_2(t)$ and $e_1(t)$. As in the case of $Z(j\omega;t)$ and $Y(j\omega;t)$, we may write

$$e_2(t)=G(p;t)e_1(t). \tag{45}$$

Formally, the expression for the gain of a variable network may be obtained in much the same manner as in the case of a fixed network. Consider for example the network shown in Fig. 1 in which N_1 and N_2 are two variable networks whose impedances are respectively, $Z_1(s;t)$ and $Z_2(s;t)$; $e_1(t)$ and $e_2(t)$ represent respectively the input and output voltages, and the problem is to find an expression for the gain $G(s;t)$ in terms of the component impedances of the network.

The input admittance of the network under consideration may be written as:

$$Y(p;t)=[Z_1(p;t)+Z_2(p;t)]^{-1} \tag{46}$$

and hence the input current $i(t)$ may be expressed as

$$i(t)=[Z_1(p;t)+Z_2(p;t)]^{-1}e_1(t). \tag{47}$$

The output voltage $e_2(t)$ is related to $i(t)$ through

$$e_2(t)=Z_2(p;t)i(t). \tag{48}$$

By combining Eqs. (48) and (47), and making use of the multiplication rule of system functions (Eq. (25)) it follows that the expression for the gain is:

$$G(s;t)=Z_2(p+s;t)[Z_1(s;t)+Z_2(s;t)]^{-1} \tag{49}$$

or

$$G=Z_2*(Z_1+Z_2)^{-1}. \tag{49a}$$

The corresponding expression for a fixed network reads:

$$G(s)=Z_2(s)[Z_1(s)+Z_2(s)]^{-1}. \tag{50}$$

A network configuration frequently used in practice consists of two or more four-terminal networks which are connected in tandem. Consider two such networks N_1 and N_2 (Fig. 2) and let their gains be respectively $G_1(s;t)$ and $G_2(s;t)$. From Eq. (25) it follows immediately that the overall gain $G(s;t)$ is given by the relation

$$G(s;t)=G_2(p+s;t)G_1(s;t). \tag{51}$$

In this relation the variable t in $G_2(p+s;t)$ should be treated as if it were a constant parameter and $G_1(s;t)$

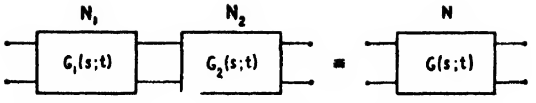


FIG. 2. Tandem connection of two variable networks.

should be treated as a function of time involving s as a parameter. It will be noticed that for the particular case of fixed networks Eq. (51) reduces to the familiar relation

$$G(s)=G_2(s)G_1(s). \tag{52}$$

As an application of Eq. (51) consider the case where N_1 is an amplitude modulator whose gain is

$$G_1(s;t)=A(t) \tag{53}$$

and N_2 is a phase modulator whose gain is expressed by

$$G_2(s;t)=e^{-s\alpha\cos\beta t} \tag{54}$$

where α and β are constants and $A(t)$ is a specified function of time.

From Eq. (51) the gain of the over-all system is found to be

$$G(s;t)=e^{-(p+s)\alpha\cos\beta t}\{A(t)\} \tag{55}$$

or

$$G(s;t)=e^{-s\alpha\cos\beta t}A(t-\alpha\cos\beta t). \tag{56}$$

It will be noted that in general an expression of the form

$$G(s;t)=G_2(s;t)G_1(t) \tag{57}$$

represents a composite network consisting of a variable network whose gain is $G_2(s;t)$ *followed* by an amplitude modulator whose gain is $G_1(t)$. Thus, Eq. (57) shows that an amplitude modulator *followed* by a phase modulator (the order of N_1 and N_2 in Fig. 2) is equivalent to the same phase modulator *followed* by an amplitude modulator whose gain is the same as that of the original modulator except for a delay in time. The same conclusion could be reached, of course, by purely physical considerations.

IV. BASIC NETWORK PROPERTIES

As was stated in the introduction, many if not most of the theorems, properties, relations, etc., that hold in the case of fixed networks, may be extended to linear varying-parameter networks. In what follows only a few such theorems, properties, relations, etc., will be considered. The topics discussed below have been chosen primarily because of their simplicity, basic importance and illustrative value.

(a) General Mesh and Node Equations of a Variable Network

The mesh and node equations of a variable network have the same form as those of a fixed network, the only difference being that all impedances and admittances in the former type of network are time-dependent quantities. Thus, for example, using operational notation the mesh equations of a variable network may be written in the following form:

$$\begin{aligned}
e_1&=Z_{11}(p;t)i_1+Z_{12}(p;t)i_2+\cdots+Z_{1n}(p;t)i_n\\
e_2&=Z_{21}(p;t)i_1+Z_{22}(p;t)i_2+\cdots+Z_{2n}(p;t)i_n\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
e_n&=Z_{n1}(p;t)i_1+Z_{n2}(p;t)i_2+\cdots+Z_{nn}(p;t)i_n.
\end{aligned} \tag{58}$$

In these equations $i_1, i_2, \cdots, i_n$ represent the mesh currents; $e_1, e_2, \cdots, e_n$ represent the loop voltages; and $Z_{11}(p;t)$, $Z_{12}(p;t)$, etc., represent the self and mutual loop impedances.

It is important to note that the principle of reciprocity does hold in the case of linear varying-parameter networks. Thus, in Eq. (58) $Z_{rs}(p;t)$ is equal to $Z_{sr}(p;t)$.

(b) Thevenin's Theorem

Thevenin's theorem may be extended to variable networks with practically no modifications in its standard form. Thus let aa' be the terminals of a variable network N (Fig. 3) and let $Z_L(s;t)$ be an impedance connected across aa'. *Then, insofar as $Z_L(s;t)$ is concerned, the network N may be replaced by a voltage source $e(t)$ in series with an impedance $Z(s;t)$, where $e(t)$ is the open-circuit voltage across aa' and $Z(s;t)$ is the impedance looking into aa' when all sources of voltage inside N are short-circuited.*

To prove the above statement we shall assume for simplicity that there is only one voltage source $e_s(t)$ inside N. Let $v(t)$ and $i(t)$ denote respectively the voltage across aa' and the current through these terminals. Because of the linearity of the system we can write

$$i(t)=Y(p;t)v(t)+Y'(p;t)e_s(t), \tag{59}$$

where $Y(p;t)$ and $Y'(p;t)$ are some as yet unspecified admittances. To obtain the physical significance of $Y(p;t)$ and $Y'(p;t)$ we proceed as follows:

First, let the voltage source $e_s(t)$ be short-circuited, i.e., $e_s(t)=0$. Then, Eq. (59) gives

$$i(t)=Y(p;t)v(t) \tag{60}$$

which shows that $Y(p;t)$ is the admittance looking into aa' when $e_s(t)$ is short-circuited.

Next, let the terminals aa' be open-circuited, i.e., $i(t)=0$, and let the voltage appearing across aa' under these conditions be denoted by $e(t)$. For this case Eq. (59) gives

$$0=Y(p;t)e(t)+Y'(p;t)e_s(t). \tag{61}$$

Combining Eqs. (59) and (61) we find

$$i(t)=Y(p;t)[v(t)-e(t)] \tag{62}$$

or, equivalently

$$v(t)-e(t)=Z(p;t)i(t) \tag{63}$$

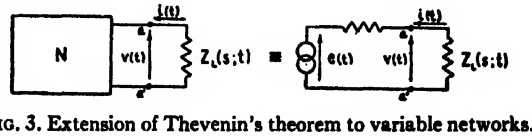


FIG. 3. Extension of Thevenin's theorem to variable networks.

where in view of Eq. (60) $Z(p;t)$ represents the impedance looking into aa' when all voltage sources in N are short-circuited. Equation (63) constitutes essentially a proof of the stated extension of Thevenin's theorem to linear varying-parameter networks.

(c) Dualization

A widely used technique in both the analysis and synthesis of fixed networks is that of dualization. This technique may also be used in variable networks in much the same manner as in the case of fixed networks. In fact, the only new aspect of dualization in the case of variable networks is that the relations between dual circuit-elements assume the following form:

Dual of a capacitance $C=f(t)$ is an inductance $L=f(t)$ and vice-versa.

Dual of resistance $R=h(t)$ is a conductance $G=h(t)$ and vice-versa.

It will be noticed that for the particular case of fixed networks the above relations reduce to the standard form. It should be emphasized that all dual properties and relations that apply to fixed networks, apply also to linear varying-parameter networks. For example, if the input impedance of a network N is $Z(p;t)$, then the input impedance of the dual of N is the inverse of $Z(p;t)$.

V. CONCLUDING REMARKS

The results obtained in the preceding sections show that the basic notion of the system function and the derived notions of impedance, admittance, gain, etc., provide the necessary basis for the extension of the conventional "fixed network" techniques of circuit analysis to linear varying-parameter networks. Much further work along these lines remains to be done, particularly in connection with the difficult problem of finding the inverse of a given system function. Nevertheless, even at its present stage of development, the circuit analysis approach outlined in this paper not only enhances understanding of the behavior of variable networks but may also be used with advantage in the solution of many problems arising in connection with the analysis and synthesis of linear varying-parameter systems.

Convection Currents in Porous Media. III. Extended Theory of the Critical Gradient

F. T. Rogers, Jr.
U. S. Naval Ordnance Test Station, Inyokern, California*

AND

H. L. Morrison
North Carolina State College of Engineering, Raleigh, North Carolina

(Received April 19, 1950)

The theory of the critical gradient for onset of thermal convection in a fluid entrapped in a porous medium is extended to allow for the exponential dependence of viscosity upon temperature and for non-linear vertical temperature-distributions which characterize transients in heat flow. Although there are approximations in the mathematical treatment, the extended theory agrees rather well with experimental data, as the simple theory does not in certain instances. New experimental data are reported for the critical gradient, obtained largely with silicone fluids in unconsolidated sands.

An earlier paper[1] under this same general title developed from basic considerations the expression

$$\beta_c=4\pi^2h^2\mu/kg\rho_0\alpha D^2, \tag{1}$$

for the (negative) mean thermal gradient which must obtain in a fluid entrapped in a porous medium in order for convection to occur. Here h^2 represents the thermal diffusivity of fluid and medium *in situ*, k the flow-permeability of the medium, μ and ρ_0 the viscosity and bottom density of the fluid, α the coefficient of cubical thermal expansion of fluid relative to medium, and g the acceleration of gravity. In the development μ was for simplicity taken to be everywhere constant, and the temperature was taken to be essentially linear in the distance above the bottom of fluid and medium.

A subsequent paper[2] reported the results of experiments designed to test the validity of Eq. (1). It was clear from these experiments that Eq. (1) is substantially correct, except for its prediction of excessive[1] β_c-values, but that its applicability is severely limited by the underlying assumptions. In the experiments μ was of necessity not constant, and in most of them convection began before a linear (i.e., steady-state) condition of heat flow could be established. The viscosity depended upon temperature according to

$$\mu\cong\mu_0\exp[-\nu(\Theta-\Theta_0)], \tag{2}$$

* Post office: China Lake, California.

[1] C. W. Horton and F. T. Rogers, Jr., J. App. Phys. 16, 367 (1945). A paper by E. R. Lapwood, Proc. Camb. Phil. Soc. 44, 605 (1948), carried out this and related developments in a somewhat more comprehensive manner.

[2] H. L. Morrison, F. T. Rogers, Jr., and C. W. Horton, J. App. Phys. 20, 1027 (1949).

where ν is a constant for a given fluid and μ_0 is the viscosity at temperature Θ_0, whereas data such as are shown in Fig. 1 imply departures of vertical temperature-distribution from linearity given by

$$\theta \cong \theta_1 \sin(\pi z/D) + \theta_2 \sin(2\pi z/D), \tag{3}$$

θ_1 and θ_2 being constants, z being distance above bottom, and D being bottom-to-top distance through the convective fluid. That the experimental results could be represented by the empirical equation,

$$\beta_{c,\,\mathrm{obs}} \cong \frac{4\pi^2 h^2 \mu_M}{kg\rho_0 D^2}\left(\frac{k\rho_0}{h^2 \mu_M K}\right)^{0.51}, \tag{4}$$

where $K = 10^{-3}$ sec.^{2}/cm^2 deg. C for c.g.s. units, bears out both points made at the beginning of this paragraph; here μ_M is the viscosity corresponding to the average temperature in the convective fluid.

THE EXTENDED THEORY

The object of extending the simple theory for β_c is to obtain an improved form of Eq. (1), which accounts for experimentally observed critical gradients as well as does Eq. (4), and perhaps to find a physical basis for Eq. (4). The process of extension is merely to allow for the sense of Eqs. (2) and (3) in a repetition of the calculations in reference 1. Thus Eq. (14) of reference 1 now appears as

$$\frac{\partial^2 w}{\partial z^2} - \nu\beta\frac{\partial w}{\partial z} - Aw + A\lambda\theta \exp[\nu(\beta z + \theta)] = 0, \tag{5}$$

where $A = l^2 + m^2$ and $\lambda = kg\rho_0\alpha/\mu_0$, μ_0 being identified as the bottom viscosity. Likewise, Eq. (15) of reference 1 is now

$$n\theta + w\beta + w\frac{\partial \theta}{\partial z} = h^2\left(\frac{\partial^2}{\partial^2 z} - l^2 - m^2\right)\theta = 0, \tag{6}$$

reflecting the fact that θ-derivatives may be comparable to β itself. Since the precision of available data does not seem sufficient to warrant retention of the second term in Eq. (3), the sense of Eq. (3) is, approximately, that $\theta = \theta_m \sin sz$; use of this in Eq. (6) leads, for small w and for $n = 0$ to

$$\theta = -\frac{w(\beta + s\theta_m)}{h^2\sigma}, \tag{7}$$

where $\sigma = s^2 + l^2 + m^2$. Then substitution of Eq. (7) into Eq. (5) yields ($\beta = -\beta_c$)

$$\frac{\partial^2 w}{\partial z^2} + \nu\beta_c\frac{\partial w}{\partial z} - Aw + \frac{A}{B}(\beta_c - s\theta_m)e^{-\nu\beta_c z}w = 0, \tag{8}$$

where $B = h^2\sigma/\lambda$, which must be solved for β_c under the boundary conditions $w(0) = w(D) = 0$.

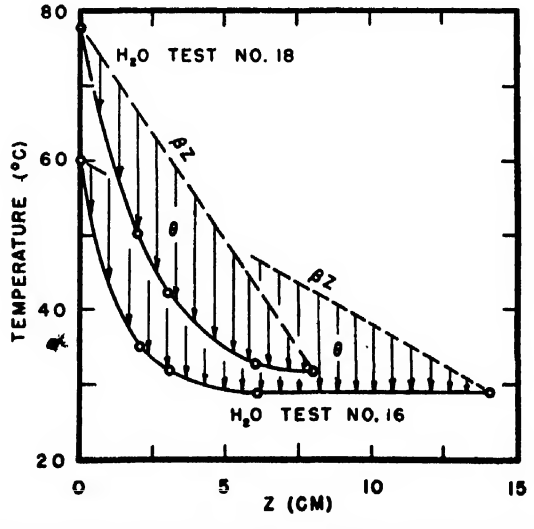


FIG. 1. Resolution of two typical non-linear vertical distributions of temperature, into linear and non-linear portions; see Eq. (3). These are the same distributions as are shown in Fig. 1 of reference 1. (Note: the axis of abscissae in Fig. 1 of reference 1 should have been labelled $D-z$ rather than z.)

Case I: $\nu = 0$

If the viscosity were actually constant, as was assumed in reference 1, but if the vertical temperature distribution were not linear, then the first derivative and the exponential in Eq. (8) would be missing. In this event Eq. (8) would have as its solution a linear combination of sine and cosine with the argument $z(A/B)^{\frac{1}{2}} \times (\beta_c - s\theta_m - B)^{\frac{1}{2}}$, and in order that boundary conditions be satisfied, this argument would equal π for $z = D$. Hence, since $A = \pi/D$ and $\sigma = 2A$ by arguments advanced in reference 1,

$$\beta_c = \frac{4\pi^2 h^2 \mu_0}{kg\rho_0\alpha D^2} + \frac{\pi\theta_m}{D}, \quad \nu = 0. \tag{9}$$

Equation (9) clearly shows how, as reference 2 reported, a negative non-linearity, θ_m, can lead to drastically smaller critical gradients than Eq. (1) would predict.

Case II: νBD small

For $\nu > 0$ but νBD small, Eq. (8) has the series solution

$$w = \sum_{j=1}^{\infty} a_j z^j,$$

where

$$\left.\begin{aligned} a_j &= \left[\frac{Q^{\frac{1}{2}(j-1)}}{(j-2)!} - \cdots\right]\frac{a_1}{j(j-1)!}, \quad j \text{ odd}, \\ a_j &= \left[\frac{Q^{\frac{1}{2}(j-2)}}{(j-2)!} - \cdots\right]\frac{\nu\beta a_1}{j(j-1)!}, \quad j \text{ even}, \end{aligned}\right\}$$

and

$$Q = A + A(s\theta_m - \beta_c)/B,$$

TABLE I. New observations of the critical gradient, with $D=10$ cm and $k=140\times10^{-8}$ c.g.s. All units are c.g.s.

Liquid	Test No.	$10^3\times\alpha$	$10^3\times h^2$	ρ	s_0 (cm)	Temperatures at onset $s=0$	2	6	10	μ_0/ρ (poises)	Observed β_c
Glycerol	5	0.495	2	1.26	3.6	100	66	44.5	38.5	0.62	6.15
Glycerol	6	0.495	2	1.26	3.6	90	60.2	36	34.8	0.98	5.52
1-DC-200	2	1.38	1.6	0.818	4.1	38.5	31	27.9	27.8	0.0074	0.93
1-DC-200	3	1.38	1.6	0.818	3.8	30	24.8	22.2	22	0.0080	0.80
1-DC-200	5	1.38	1.6	0.818	3.3	30	26	24	23.7	0.0080	0.63
10-DC-200	1	1.12	1.7	0.940	4.0	67	44	31	29.6	0.080	3.1
10-DC-200	3	1.12	1.7	0.940	4.1	70	47.6	34.2	32.8	0.078	3.72
10-DC-200	5	1.12	1.7	0.940	4.1	50	38	29	27.3	0.085	2.27
100-DC-200	1	0.923	1.7	0.968	3.0	90	56	34.4	32	0.67	5.80
100-DC-200	4	0.923	1.7	0.968	3.1	85	50	28	28	0.74	5.40
100-DC-200	6	0.923	1.7	0.968	3.3	100	60	34.5	33.6	0.675	6.60

to the first order in νBD. By suitable manipulation, this can be rearranged, so that the boundary condition $w(D)=0$ leads to

$$\pi[(\beta_c/B)-(s\theta_m/B)-1]^{\frac{1}{2}}\sin\pi[(\beta_c/B)-(s\theta_m/B)-1]^{\frac{1}{2}} -\nu\beta_c D\cos\pi[(\beta_c/B)-(s\theta_m/B)-1]^{\frac{1}{2}} +\text{higher-order terms in } \nu BD=0. \quad (10)$$

The first two terms in Eq. (10) suffice to determine β_c for the present case; since $\beta_c/B-s\theta_m/B$ must be close to 2, some reduction yields

$$\beta_c=\frac{\dfrac{4\pi^2h^2\mu_0}{kg\rho_0\alpha D^2}+\dfrac{\pi\theta_m}{D}}{1+\dfrac{2\nu D}{\pi^2}\dfrac{4\pi^2h^2\mu_0}{kg\rho_0\alpha D^2}}, \quad \nu BD \text{ small.} \quad (11)$$

The outstanding feature of Eq. (11) is its insistence that a non-constant viscosity should reduce the critical gradient.

Case III: νBD large

Equation (8) can be put into the form of Bessel's equation[3] by the transformations $\eta=w\exp(\nu\beta_c z)$ and $\zeta=[4A(\beta_c/B-s\theta_m/B)/\nu^2\beta_c^2]^{\frac{1}{2}}\exp(-\frac{1}{2}\nu\beta_c z)$,

$$\eta''+\frac{\eta'}{\zeta^2}+\left[1-\frac{1+(4A/\nu^2\beta_c^2)}{\zeta^2}\right]=0, \quad (12)$$

and this is well suited to the study of less simple convective environments. In terms of

$$q=(1+4A/\nu^2\beta_c^2)^{\frac{1}{2}},\quad \zeta_0=2\pi(\beta_c/B-s\theta_m/B)^{\frac{1}{2}}/\nu\beta_c D,\quad \text{and}\quad \zeta_D=\zeta_0\exp(-\tfrac{1}{2}\nu\beta_c D),$$

its solution, a linear combination of Bessel functions of

TABLE II. (Negative) critical gradients calculated by various forms of the theory.

Liquid	Test No.	μ_0 (poises)	$2B$ (°C/cm)	ν (°C^{-1})	θ_m (°C)	β_c, Critical gradients at onset (°C/cm) Observed	Calculated by Eq. (1)	Eq. (4)	Eq. (11)	Eq. (14)
Glycerol	2	0.0012	1.60	0.0860	40	9.4	550	14	—	6.5
Glycerol	5	0.0148	13.76	0.0860	25	6.15	572	11	4.76	(1.76)
Glycerol	6	0.0316	29.2	0.0860	27	5.52	904	14	13.75	(1.23)
CCl_4	3	0.0084	2.77	0.0126	0	0.8	2.86	1.1	2.58	(0)
Water	1	0.00448	16.34	0.0142	13	3.8	26.3	3.1	8.33	—
Water	5	0.00303	19.72	0.0142	20	6.1	34.1	3.5	8.57	—
Water	6	0.00385	19.94	0.0142	18	5.4	38.4	3.7	9.08	—
Water	7	0.00436	15.88	0.0142	14	3.95	26.3	3.1	7.88	—
Water	8	0.00436	15.88	0.0142	15	4.1	27.0	3.1	7.63	—
Water	9	0.00469	8.72	0.0142	17	2.2	14.1	1.6	3.62	—
Water	11	0.00317	18.06	0.0142	25	6.9	33.0	4.4	5.81	—
Water	13	0.0028	436	0.0142	—	>7.5	460	<12	<32	(>0?)
Water	15	0.00412	15.00	0.0142	17	3.75	24.3	3.0	6.75	—
Water	16	0.00468	8.72	0.0142	15	2.2	14.1	1.6	3.96	(6.25)
Water	18	0.00368	20.94	0.0142	20	5.6	40.2	4.7	8.82	(10.79)
1-DC-200	2	0.0072	2.94	0.0120	6.5	0.93	3.02	1.1	0.84	(13.88)
1-DC-200	3	0.0080	3.27	0.0120	4	0.80	3.26	1.1	1.84	—
1-DC-200	5	0.0080	3.26	0.0120	2.5	0.63	3.26	1.1	2.27	—
10-DC-200	1	0.0475	22.08	0.0152	23	3.1	36.4	4.7	(8.85)	7.50
10-DC-200	3	0.0469	21.80	0.0152	19	3.72	40.0	4.5	(9.46)	6.45
10-DC-200	5	0.0601	27.94	0.0152	9.5	2.27	39.1	4.5	(13.43)	4.30
100-DC-200	1	0.322	176.6	0.0177	26.5	5.80	367	16.4	(22.94)	2.12
100-DC-200	4	0.362	198.6	0.0177	30	5.40	407	17.1	(23.58)	2.46
100-DC-200	6	0.287	157.4	0.0177	33	6.60	370	17.6	(22.39)	2.90

[3] H. Massey and C. Mohr, Proc. Roy. Soc. London A152, 693 (1935).

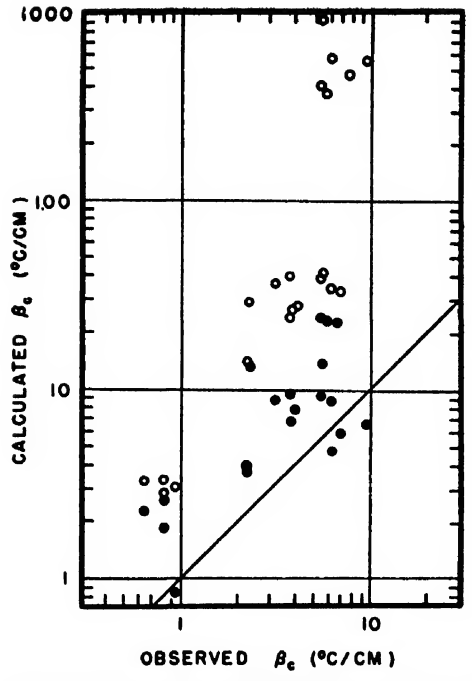


FIG. 2. Comparison of the simple theory, Eq. (1) (open circles), and the extended theory, Eqs. (11) and (14) (filled circles), with observational data. Points on the straight line would represent perfect agreement. The general tendency of points to lie above the line was anticipated in reference 1, and is attributable in part to a choice of boundary conditions differing somewhat from those which obtained in the experiments.

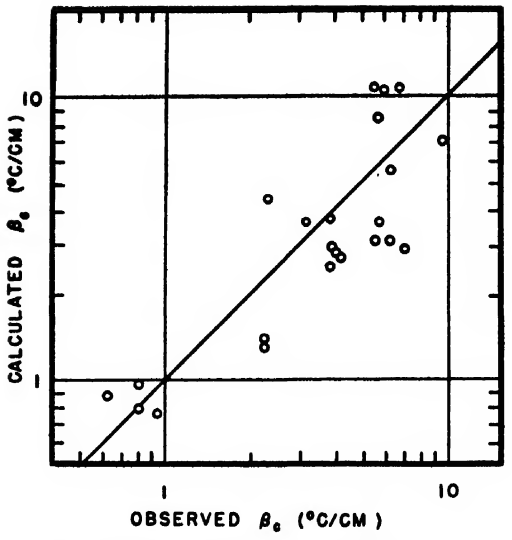


FIG. 3. Comparison of the improved empirical formula, Eq. (15), with observational data. Points on the straight line represent perfect agreement. Although Eq. (15) predicts critical gradients which depart in some instances by a factor of three from observed values, it is perhaps the best formula available at the present time.

the first and second kinds of order q, can be written as

$$J_q(\zeta_0)J_{-q}(\zeta_D)-J_q(\zeta_D)J_{-q}(\zeta_0)=0, \qquad (13)$$

and in this form satisfies the condition $w(D)=0$. Since the domain of large $\nu\beta_c D$-values is not of great interest at present, it suffices to indicate what Eq. (13) predicts for the typical value $\nu\beta_c D=5.62$ (i.e., $q=\frac{3}{2}$); in this instance Eq. (13) reduces to

$$\zeta_0\cong 2^{\frac{1}{2}}\pi,$$

whence for θ_m sufficiently negative

$$\beta_c|_{\theta_m\ll 0}\cong\frac{1}{\nu D}\left(\frac{-4\pi\theta_m/D}{4\pi^2h^2\mu_0/kg\rho_0\alpha D^2}\right)^{\frac{1}{3}}, \quad \nu\beta_c D=5.62. \qquad (14)$$

Computational experience suggests that Eq. (14) holds as a solution of Eq. (13), in an approximate sense for $3<\nu\beta_c D<9$.

COMPARISON WITH EXPERIMENT

For the purpose of providing a larger body of data against which to compare the above extended theory, additional experiments have been carried out by the technique reported in reference 2. In most of these experiments silicone fluids were used, each being identified by type (Dow-Corning DC-200) and (by a suitable prefix) by its nominal kinematic viscosity in centistokes at 25°C. Ciba oil red C Y dye was employed[4] in the silicone experiments as a visible indicator of the state of fluid motion, and an unidentified blue dye in the few glycerol experiments. Table I exhibits typical data obtained from these new experiments.

Table II contains the observed mean critical gradients for all tests reported in reference 2 as well as for those in Table I, together with gradients calculated by the various formulas of theory.

It is obvious by inspection of Table II that the extended theory as embodied in Eqs. (11) and (14), is in far better accord with the observational data than is the original simple theory summarized by Eq. (1). Figure 2 illustrates this, and at the same time demonstrates the quasi-statistical character of the accordance. For the present it does not seem worth while to expend much effort on mathematical refinements of the extended theory, because many components of experimental technique deserve prior improvement.[5]

While the entries in Table II show that the empirical Eq. (4) is in about as good agreement with observation as are Eqs. (11) and (14), a brief search has been made for a better empirical law. It has been found, as Fig. 3 shows graphically, that a more precise relation is

$$\beta_{c(\text{empir})}=0.415(\beta_{c(\text{Eq. (4)})}+0.455\beta_{c(\text{Eqs. (11), (14)})}), \qquad (15)$$

the notation here being self-explanatory. Equation (15) predicts critical gradients differing from the observed ones by not more than a factor of three, and so is probably the most accurate formula available at the present time.

[4] The authors are indebted to the Ciba Company, Inc., New York, New York, for the gift of oil red C Y dye used in these experiments.

[5] See, e.g., "$\partial T/\partial t$ at the Onset of Convection in Porous Media," by F. T. Rogers, Jr., and H. L. Morrison, read before the 1950 meeting of the Southeastern Section of the American Physical Society in Baton Rouge.

Theory of Axially Slitted Circular and Elliptic Cylinder Antennas

D. R. RHODES
Department of Electrical Engineering, The Ohio State University Research Foundation, Columbus, Ohio
(Received April 20, 1950)

A method proposed by Sommerfeld for solving boundary value problems involving discontinous surfaces has been applied to the general case of a plane electromagnetic wave of arbitrary direction of incidence and polarization diffracting about one or more sections of perfecting conducting circular and elliptic cylinders of infinite length. The solution is in series form, where the series coefficients are independent for the special case of slots of infinitesimal width (slits). Radiation from the slitted cylinder is restricted to discrete right circular cones about the cylinder axis, each cone corresponding to an ordinary wave-guide mode in the cylinder, while slots of finite width radiate over a continuous range of conical angles. For slots of small but finite width, the relative pattern in the cone $\alpha=\alpha_0$ about a circular cylinder of radius a is the same as in a plane normal to the axis of a cylinder of radius $a\sin\alpha_0$; patterns of the slotted elliptic cylinder are similarly related, where the analog of radius is distance between foci. Conical patterns are shown for the principal *TE* and *TM* wave-guide modes in circular cylinders containing one and two diametrically opposed axial slits.

A NUMBER of papers have appeared in the past ten years in which controllable radiation from long axial slots cut in the walls of rectangular wave guide has been reported. In connection with an experimental investigation of radiation from slotted cylinder traveling wave antennas conducted at The Ohio State University[1] it was apparent that a theoretical analysis of this type of antenna would be desirable. A rigorous boundary-value solution involving no assumptions that limit the validity of the solution for a long slot in a rectangular wave guide is, in general, a difficult problem. A slotted cylindrical wave guide whose surface is defined by a single constant coordinate, however, can be treated as a diffraction problem using a method proposed by Sommerfeld[2] in connection with scattering by a cylindrical mirror. Although Sommerfeld applied the method only to two-dimensional scalar fields incident normally on a section of a circular cylinder, the method is easily extended to include the general vector field incident at any angle on elliptic as well as circular cylinders of one or more sections. The solution involves rather lengthy numerical computations in general, but a very simple solution results when the slots are assumed to be narrow. It is the special case of slots of infinite length and vanishing width, or "slits" as they shall be referred to,[3] in a perfectly conducting circular and elliptic cylindrical shell of zero wall thickness which is treated in this paper.

1. DIFFRACTION AROUND SECTIONS OF CIRCULAR CYLINDRICAL SHELLS

The coordinate system of the slotted cylinder is shown in Fig. 1. The fields diffracted about the cylinder by a plane wave incident at an arbitrary angle with respect to the cylinder axis can be obtained by expressing the electric and magnetic fields in terms of two independent single component Hertzian vectors directed along the z axis.[4] Each of the two scalar functions are then solutions of the ordinary scalar wave equation. The required solution is that linear combination of characteristic functions whose coefficients are such that the boundary conditions at the cylindrical surface and at infinity are satisfied.

A single section and two equal diametrically opposed sections of a circular cylinder are shown in cross section

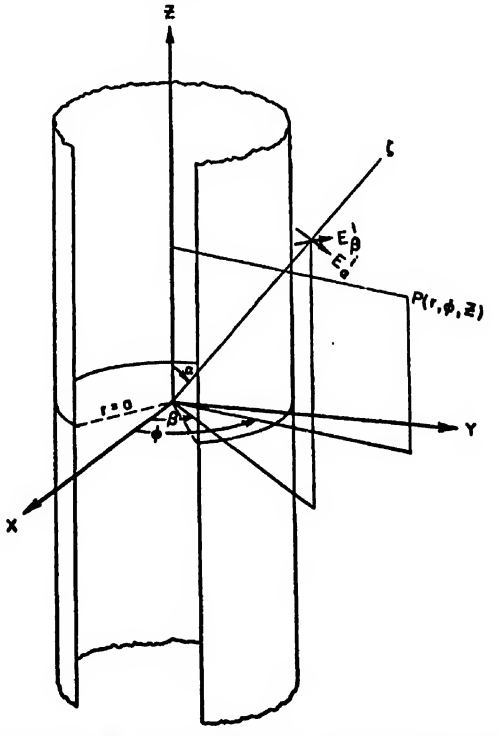


FIG. 1. Coordinate system defining the spherical angles of arrival α and β of the incident plane wave along the ζ-axis and the cylindrical coordinates r, φ, z of an arbitrary point in space.

[1] Hines, Krausz, Rhodes, Rumsey, and Walter, Proc. I.R.E. (to be published).
[2] A. Sommerfeld, *Partial Differential Equations* (Academic Press Inc., New York, 1949), pp. 29–31 and pp. 159–162.
[3] The term "slit" carries the connotation of narrowness to a greater degree than the term "slot" and hence shall be used to specify a slot in this more restrictive sense.
[4] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), pp. 349–51.

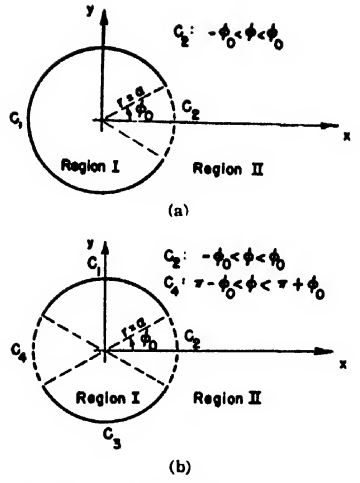


FIG. 2. Cross section of an infinitely long circular cylinder of radius a with (a) a single conducting section, and (b) two equal diametrically opposed conducting sections.

in Fig. 2. These two cases only will be considered. Boundary conditions at the cylindrical surface require that the tangential components of electric field and the normal components of magnetic field vanish over the conducting sections, and that all components of electric and magnetic field match over the aperture separating the interior and exterior regions of the cylinder. The requirement that the scattered radiation consist of outward traveling waves at infinity is satisfied by choosing Hankel functions of the second kind as the radial function in the scattered waves.

For any arbitrary polarization of the incident plane wave it is necessary to use both Hertz vectors. However, if the incident wave is resolved into two orthogonal components, one in and one normal to the plane of incidence, the diffracted field produced by each component can be independently determined using only one Hertz vector. The total field is obtained by superposing the two at every point in space. Elliptically as well as linearly polarized incident waves can be used by including the appropriate time phase between the two orthogonal components.

The incident electric field will, in general, have components $E_\alpha{}^i$ and $E_\beta{}^i$. Each of these components is of the form[5]

$$E^i = E_0 e^{ikz\cos\alpha+i\omega t}\sum_{-\infty}^{\infty} i^n J_n(kr\sin\alpha)e^{in(\varphi-\beta)}. \quad (1)$$

Considering first the component $E_\alpha{}^i$, the fields are derivable from a transverse magnetic Hertzian potential Π_z proportional to E_z. Resolving $E_\alpha{}^i$ into its cylindrical components, the incident potential $\Pi_z{}^i$ is found from $E_z{}^i$ to be

$$\Pi_z{}^i = K\sum_{n=-\infty}^{\infty} i^n J_n(kr\sin\alpha)e^{in(\varphi-\beta)}, \quad (2)$$

where

$$K = -\frac{E_{0\alpha}}{k^2\sin\alpha}e^{ikz\cos\alpha+i\omega t}. \quad (3)$$

Designating the potential inside the cylinder by $\Pi_z{}^{\mathrm{I}}$ and the potential outside the cylinder by $\Pi_z{}^{\mathrm{II}}$, the latter of which can be written as the sum of the incident and scattered potentials

$$\Pi_z{}^{\mathrm{II}} = \Pi_z{}^{\mathrm{I}} + \Pi_z{}^s, \quad (4)$$

the potentials can be written in the following form:

$$\Pi_z{}^{\mathrm{I}} = K\sum_{n=-N}^{N} a_n J_n(kr\sin\alpha)e^{in(\varphi-\beta)}, \quad (5)$$

$$\Pi_z{}^{\mathrm{II}} = K\sum_{n=-N}^{N}[i^n J_n(kr\sin\alpha) + c_n H_n{}^{(2)}(kr\sin\alpha)]e^{in(\varphi-\beta)}, \quad (6)$$

where the summations are limited to a finite number of terms. The exact solution can be approximated as closely as desired by series containing $2N+1$ terms if the arbitrary positive integer N is chosen sufficiently large.

The boundary conditions at $r=a$ must determine the unknown coefficients a_n and c_n uniquely. Three necessary and sufficient independent boundary conditions on the potentials are the following:

$$(\Pi_z{}^{\mathrm{I}} = \Pi_z{}^{\mathrm{II}})c_1 + c_2, \quad (7a)$$

$$(\Pi_z{}^{\mathrm{I}} = \Pi_z{}^{\mathrm{II}} = 0)c_1, \quad (7b)$$

$$\left(\frac{\partial\Pi_z{}^{\mathrm{I}}}{\partial r} = \frac{\partial\Pi_z{}^{\mathrm{II}}}{\partial r}\right)_{c_2}. \quad (7c)$$

Since the potentials in the two regions must match over the periodic orthogonal interval $0\leq\varphi\leq 2\pi$, by Eq. (7a), the coefficients may be equated termwise:[6]

$$a_n J_n = i^n J_n + c_n H_n{}^{(2)}. \quad (8)$$

Boundary condition Eq. (7b) requires that

$$\left(\sum_{n=-N}^{N} a_n J_n e^{in(\varphi-\beta)} = 0\right)c_1, \quad (9)$$

and boundary condition Eq. (7c) together with Eq. (8) relating a_n and c_n requires that

$$\left(\sum_{n=-N}^{N}\frac{a_n - i^n}{H_n{}^{(2)}}e^{in(\varphi-\beta)} = 0\right)_{c_2}. \quad (10)$$

Equations (9) and (10) relate the coefficients a_n over the two subintervals C_1 and C_2. The solution for the coefficients is referred to by Sommerfeld[3] as a "non-final determination of coefficients," that is, each coefficient depends upon every other coefficient in the series because of the non-orthogonal subintervals. This is a very

[5] Reference 4, p. 371–2. The notation used is essentially that of Stratton with the exception that the time convention has been chosen to be $e^{i\omega t}$ rather than $e^{-i\omega t}$.

[6] The convention will be adopted that all solutions of Bessel's differential equation evaluated at $r=a$ will be written without the argument.

essential distinction from the problem of diffraction around a completely closed conducting surface in which each coefficient is uniquely determined independently of the others. The method of solution proposed by Sommerfeld is that of minimizing the total mean square error incurred over the two subintervals, thereby obtaining the best approximation possible for the finite number of terms used in the series. Since it is required that each summation in Eqs. (9) and (10) is to approximate the null function, the total mean square error over C_1 and C_2 is given by

$$M=\int_{C_1}\left[\sum_{n=-N}^{N} a_n J_n e^{in(\varphi-\beta)}\right]^2 d\varphi + \int_{C_2}\left[\sum_{n=-N}^{N}\frac{a_n-i^n}{H_n^{(2)}} e^{in(\varphi-\beta)}\right]^2 d\varphi. \quad (11)$$

Differentiating M with respect to each unknown coefficient results in the following set of $2N+1$ equations, which must be satisfied simultaneously for the $2N+1$ unknown a_n's:

$$2\pi a_p J_p^2 e^{-ip\beta} - \sum_{n=-N}^{N}\left[J_n J_p - \frac{1}{H_n^{(2)}H_p^{(2)}}\right]\mathfrak{a}_{np} e^{-in\beta} a_n = \sum_{n=-N}^{N}\frac{i^n e^{-in\beta}}{H_n^{(2)}H_p^{(2)}}\mathfrak{a}_{np}, \quad |p|\leq N, \quad (12)$$

where δ_{np} is Kronecker's delta and where the integral of products of angular functions over C_2 is given by

$$\mathfrak{a}_{np}=\frac{2\sin(n-p)\varphi_0}{n-p}, \quad (13)$$

$\mathfrak{a}_{pp}$ being defined as the limit when $n\to p$.

The analysis for transverse electric fields excited by the other component E_β^i of the incident plane wave is exactly analogous to the preceding analysis for transverse magnetic fields, where K is replaced by

$$K^*=-\frac{E_{0\beta}}{\eta k^2\sin\alpha} e^{ikx\cos\alpha+i\omega t},$$

where $\eta=(\mu/\epsilon)^{\frac{1}{2}}$. The equations relating coefficients a_n^* and c_n^* correspond to those relating a_n and c_n in Eqs. (8) and (12) when the radial functions are replaced by their derivatives.

The fields diffracted around a multiple section cylindrical shell by an incident plane wave are treated in the same manner as the single section cylinder. As with the single-section cylinder the boundary conditions must be satisfied over each of the conducting sections and apertures between sections. Minimizing the total mean square error over the cylinder again gives the best approximation in the sense of least squares. Although the method is applied here only to the case of two equal diametrically opposed sections, it can be easily extended to include more than two sections.

For the transverse magnetic component E_α^i of a plane wave incident on the cylinder shown in cross section in Fig. 2b, boundary conditions again require that a_n and c_n be related by Eq. (8) and that Eqs. (9) and (10) must hold over the subintervals C_1, C_3 and C_2, C_4, respectively. The resulting relationship between coefficients a_n then reduces to the following set of equations:

$$\pi a_p J_p^2 e^{-ip\beta} - \sum_{n=-N}^{N}{}'\left[J_n J_p - \frac{1}{H_n^{(2)}H_p^{(2)}}\right]\mathfrak{a}_{np} e^{-in\beta} a_n = \sum_{n=-N}^{N}{}'\frac{i^n e^{-in\beta}}{H_n^{(2)}H_p^{(2)}}\mathfrak{a}_{np}, \quad |p|\leq N, \quad (14)$$

where the primed summation is on even values of n when p is even and odd values of n when p is odd. The set of equations relating coefficients a_n^* for the transverse electric component E_β^i correspond to Eqs. (14) when the radial functions are replaced by their derivatives.

It can be seen from Eq. (12) that the coefficients a_n are functions of the angles α and β of the incoming plane wave. This is a severe limitation when solving for the pattern of the scattered field or of a detecting probe near the conducting surface, since it is necessary to repeat the solution of the system of equations for each angle of incidence. The limitation vanishes, however, for the special case of a vanishingly narrow slot in a cylinder.

2. SERIES COEFFICIENTS FOR NARROW SLOTS

As the slot width vanishes, the set of Eqs. (12) relating coefficients a_n reduces to a set of independent equations, each equation involving only a single coefficient, since C_1 approaches the interval of orthogonality of the angular functions. When the angular slot width $2\varphi_0$ is small, the integrals over C_1 are essentially independent of n and p, reducing to

$$\mathfrak{a}_{np}\approx 2\varphi_0. \quad (15)$$

Consider first the transverse magnetic component of the plane wave incident on the cylinder. When the argument $ka\sin\alpha$ is not a zero of J_q, then by choosing φ_0 vanishingly small all of the terms in the summation on a_n become vanishingly small with respect to the unsummed term containing a_p, Eqs. (12) then becoming

$$2\pi a_p J_p^2 e^{-ip\beta}=\frac{2\varphi_0}{H_p^{(2)}}F_c(\alpha,\beta), \quad |p|\leq N, \quad (16)$$

where

$$F_c(\alpha,\beta)=\sum_{n=-N}^{N}\frac{i^n e^{-in\beta}}{H_n^{(2)}}. \quad (17)$$

When $ka\sin\alpha$ is a zero of J_q, then for $p\neq q$ the set of Eqs. (12) reduce as before to Eq. (16). For $p=q$, however, the unsummed term containing a_q is zero and all

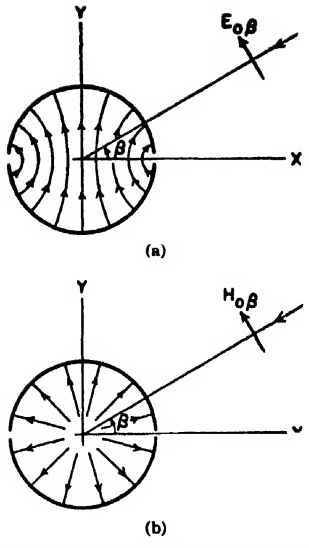


FIG. 3. Electric field distribution in the xy-plane within a slitted cylinder excited by a plane wave incident at (a) α_{11}', and (b) α_{01}.

of the a_n's in the summation are vanishingly small except the one containing a_q; hence at $p=q$ the set of Eqs. (12) reduces to

$$a_q e^{-iq\beta} = -iN_q F_c(\alpha, \beta), \quad |q| \leq N, \tag{18}$$

which determines each coefficient a_q explicitly. The coefficients a_n^* for transverse electric excitation correspond as before when the radial functions are replaced by their derivatives. A similar solution follows immediately from Eq. (14) for the double section cylinder.

The scattered radiation reduces to that of a completely closed cylinder in the limit as φ_0 vanishes, since a_n vanishes everywhere except at the zeros of J_q. For a_n to vanish everywhere including the zeros of J_q it is only necessary to let φ_0 vanish before J_q passes through its zeros. The coefficients for the scattered radiation are then obtained directly from Eq. (8).

3. MODES IN THE SLITTED CIRCULAR CYLINDER

a. Existence of Wave-Guide Modes at Certain Critical Incident Angles

The electric and magnetic fields everywhere inside a cylinder containing one or more slits have vanishingly small amplitudes due to the factor φ_0, except when the incident plane wave arrives at such an angle that

$$J_q(ka \sin\alpha_{qm}) = 0, \tag{19a}$$

or

$$J_q'(ka \sin\alpha_{qm}') = 0, \tag{19b}$$

where α_{qm} and α_{qm}' denote the incident angles for which the mth zeros of J_q and J_q' occur, respectively. The series representing each field component then reduces to two terms, namely, those for which $n=\pm q$. At α_{qm} the components of electric field in a cylinder containing a single slit reduce to the following:

$$E_r^{(1)} = -\epsilon_q E_{0\alpha}[\cos\alpha_{qm} J_q'(kr \sin\alpha_{qm}) \times \cos q\varphi] N_q F_c(\alpha, \beta), \tag{20a}$$

$$E_\varphi^{(1)} = \epsilon_q E_{0\alpha}\left[\cos\alpha_{qm} \frac{q}{kr \sin\alpha_{qm}} \times J_q(kr \sin\alpha_{qm}) \sin q\varphi\right] N_q F_c(\alpha, \beta). \tag{20b}$$

$$E_z^{(1)} = i\epsilon_q E_{0\alpha}[\sin\alpha_{qm} J_q(kr \sin\alpha_{qm}) \times \sin q\varphi] N_q F_c(\alpha, \beta), \tag{20c}$$

where ϵ_q is Newman's number; the common factor $\exp(ikz \cos\alpha_{qm} + i\omega t)$ has been omitted. Equations (20) are precisely the expressions given by Stratton[7] for the electric field components of the TM_{qm} mode in a hollow circular cylinder wave guide if Stratton's coefficients a_{qm} are chosen to be

$$a_{qm} = i\epsilon_q E_{0\alpha}(\lambda/\lambda_{qm}) N_q F_c(\alpha, \beta),$$

and if the critical angle α_{qm} is related to the cut-off wave-length λ_{qm} of the TM_{qm} circular wave-guide mode in such a way that

$$\sin\alpha_{qm} = \lambda/\lambda_{qm} = u_{qm}/ka, \tag{21}$$

where u_{qm} is the mth zero of J_q. Similarly, it can be shown that a plane wave incident at $\alpha=\alpha_{qm}'$ excites the TE_{qm} mode in a slitted circular cylinder, where α_{qm}' is related to the wave-length of the exciting plane wave and to its cut-off wave-length by Eq. (21). The field distribution of the lowest order TE and TM modes excited in a slitted cylinder are shown in Fig. 3.

An interesting philosophical but physically useless observation is the fact that a non-vanishing energy bearing field can exist in the cylinder in the ideal case of a plane wave incident at the critical conical angles on a perfectly conducting cylindrical shell, even when the shell is completely closed. These fields are physically unobtainable, since they must collapse if any energy is withdrawn from the system.

b. Modes Dependent on Slit Width

At all angles of incidence other than those in the cones α_{qm} and α_{qm}' the coefficients of the series for the fields in the cylinder are proportional to the slit width. Even at the critical angles defined by these cones all but one pair of coefficients are proportional to slit width, the fields associated with this pair being identified with ordinary wave-guide modes in the cylinder. As a consequence, for incidence at angles other than the critical angles all fields in the cylinder vanish with slit width. Since there has been some interest expressed by various

[7] Reference 4, p. 541, Eqs. (75). A typographical error is apparent in the angular function in the expression for E_z.

authors[8-10] in one special case in which these vanishing modes are excited, namely, the case of a plane wave incident normally on a slitted cylinder of arbitrary radius, these modes will be briefly considered.

For a single slit in a cylinder, the axial component of electric field for TM waves and the angular component of electric field for TE waves, for incidence at angles other than α_{qm} and α_{qm}', are given by the following:

$$E_z^{(1)} = -\frac{\varphi_0}{\pi} E_{0\alpha} \sum_{p=0}^{N} \frac{\epsilon_p}{J_p^2 H_p^{(2)}} \times [\sin\alpha J_p(kr\sin\alpha)\cos p\varphi] F_c(\alpha, \beta), \quad (22a)$$

$$E_\varphi^{(2)} = -i\frac{\varphi_0}{\pi} E_{0\beta} \sum_{p=0}^{N} \frac{\epsilon_p}{J_p'^2 H_p^{(2)\prime}} \times [J_p'(kr\sin\alpha)\cos p\varphi] F_c'(\alpha, \beta), \quad (22b)$$

where $F_c'(\alpha, \beta)$ corresponds to $F_c(\alpha, \beta)$ when the radial functions are replaced by their derivatives. For a cylinder containing two diametrically opposed slits the conical pattern factors $F_c(\alpha, \beta)$ and $F_c'(\alpha, \beta)$ are each summed over even values of n when p is even and odd values of n when p is odd. Individual terms in the summation on p in Eqs. (22) do not represent ordinary wave-guide modes, since no relation such as that given by Eq. (21) can exist for arbitrary values of α.

At normal incidence, Eqs. (22) contain the same conical pattern factors which Sinclair[8] obtained in a plane normal to the cylinder axis for a very narrow axial slot in a cylinder when considered as a transmitting antenna. Sinclair resolved an assumed uniformly distributed tangential electric field at $r=a$ into a Fourier series and equated this series termwise to a second series representing the total electric field for $r \geq a$. Using Sinclair's method, Papas and King[9] computed the current distribution on the surface of a circular cylinder excited by an electric field across an axial slit. Sinclair has also shown[10] that the pattern, in a plane normal to the cylinder axis, of arrays of axial slits on the surface of conducting circular cylinders can be calculated by superposing the fields caused by each slit. The patterns of two diametrically opposed slits obtained by using Sinclair's method of arraying slits are the same as given here.

4. OPERATION AS A TRAVELING WAVE ANTENNA

Operation of the slitted cylinder as a receiving antenna is readily apparent. As the conical angle α is continuously varied, the fields in the cylinder abruptly change at α_{qm} and α_{qm}' from fields having vanishingly small amplitudes (proportional to the slit width) into a single TM or TE wave-guide mode whose amplitude is independent of the slit width. Since the zeros of J_q and J_q' are distinct, no two wave-guide modes are excited at the same angle of incidence and hence each mode is distinct (exceptions are the TE_{0m} and TM_{1m} modes, which are excited at the same incident angle since $J_0' = -J_1$). A receiving probe placed in the cylinder has an open-circuit voltage induced in it proportional to the field at the probe. A slotted cylinder antenna should exhibit directional characteristics dependent on slot width, approaching infinite directivity as the slot width vanishes.

The critical conical angles were determined several years ago by Hansen[11] by considering axially slitted wave guides as transmitting traveling wave antennas. Maximum radiation occurs at those angles for which radiation from every infinitesimal element of the slit aperture interferes constructively. Referring to Fig. 4, maximum constructive interference occurs when α_{qm} or α_{qm}' is related to the excitation wave-length and to the guide wave-length by

$$\cos\alpha_{qm} = \lambda/\lambda_g. \quad (23)$$

Equation (23), obtained from simple interference considerations, is equivalent to Eq. (21) obtained by rigorous solution of the boundary value problem.

The critical cones associated with each wave-guide mode are distinguished by the roots u_{qm} and u_{qm}'. For any given wave-guide mode these cones are related to the electrical radius of the cylinder by Eq. (21). The angles α_{qm} and α_{qm}' are plotted in Fig. 5 as a function of the electrical radius a/λ for cylinders not exceeding a radius of one wave-length. The critical cones asymptoti-

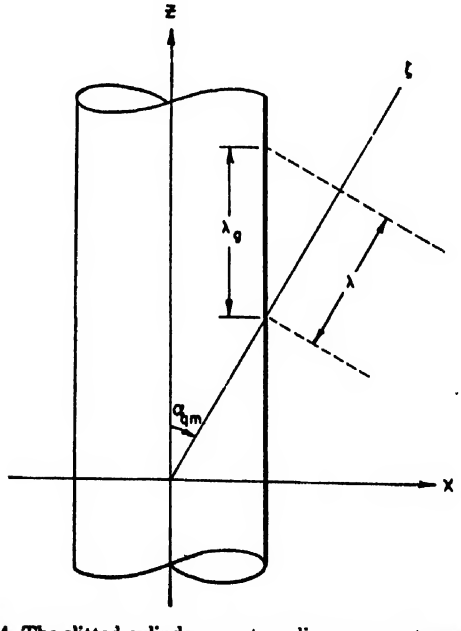


FIG. 4. The slitted cylinder as a traveling wave antenna. Maximum constructive interference occurs in cones defined by $\cos\alpha_{qm} = \lambda/\lambda_g$, where λ is free space wave-length and λ_g is guide wave-length.

[8] Sinclair, Jordan, and Vaughn, Proc. I.R.E. 35, 1451–62 (1947).
[9] C. H. Papas and R. King, Quart. App. Math. 7, 175–82 (1949).
[10] George Sinclair, Proc. I.R.E. 36, 1487–92 (1948).
[11] W. W. Hansen, U. S. Patent No. 2402622.

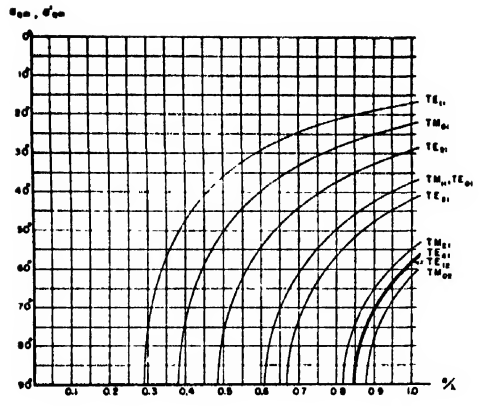


FIG. 5. Critical conical angles α_{qm} and α_{qm}' as a function of electrical radius of the slitted circular cylinder.

cally approach the cylinder axis as the excitation frequency is increased. In contrast to the unbounded upper frequency limit, each mode has a sharply defined low frequency limit. The low frequency limit occurs when maximum radiation is in the plane normal to the cylinder axis, and corresponds to the condition of cut-off in the cylindrical wave guide. For all frequencies below the low frequency limit the wave guide is operating below its cut-off frequency, and the energy propagating inside the cylinder is rapidly attenuated.

The slitted cylinder antenna is fundamentally linearly polarized, the electric field radiated by TM modes being E_α and that by TE modes being E_β.

The theory of the slitted cylinder can be used to predict the conical pattern and critical cones of wave-guide-fed slits in cylinders. The critical cones are determined by the phase velocity of waves traveling in the slit, which is determined entirely by the wave-guide cross-section; the conical pattern depends on the cross sections of both wave guide and cylinder. For the case of a circular wave guide feeding a slotted circular cylinder as shown in Fig. 6a, the argument of the radial functions in the conical pattern factor $F_c(\alpha_{qm}, \beta)$ for the TM_{qm} mode in the wave-guide is

$$kb \sin\alpha_{qm} = (b/a)u_{qm}. \tag{24}$$

Two or more wave-guide-fed slits in a cylinder, as suggested in Fig. 6b, may be independently excited and the radiation pattern calculated by Sinclair's method of superposing fields.

5. RADIATION PATTERNS

The relative conical radiation patterns of the slitted cylinder are plots of the conical pattern factors $F_c(\alpha, \beta)$ and $F_c'(\alpha, \beta)$. The pattern factors are complex functions of β from which the amplitude and phase patterns can be obtained. The conical pattern is fixed for any given wave-guide mode in a slitted cylinder regardless of the cylinder radius or critical conical angle. In Fig. 7 is shown the conical amplitude patterns characteristic of the lowest order TE and TM modes in a circular cylinder containing one or two diametrically opposed slits. An example of the three dimensional pattern of a slitted cylinder is shown in Fig. 8 for a $\frac{3}{8}\lambda$ diameter cylinder containing one or two diametrically opposed slits. The patterns in the critical cones are those shown in Fig. 7 when plotted on an absolute scale. The critical cones for the cylinder diameter chosen are $\alpha_{11}' = 42°$ and $\alpha_{01} = 61°$.

Since the incident angle α appears in the pattern factor only in the product $ka \sin\alpha$, it follows that the relative pattern in any cone $\alpha = \alpha_0$ about a cylinder of radius $r = a$ is the same as the relative pattern in a plane normal to the axis of a cylinder of radius $r = a \sin\alpha_0$.

6. SLITTED ELLIPTIC CYLINDER

The preceding method outlined for the slitted circular cylinder can be applied equally well to the slitted elliptic cylinder, the only other closed cylindrical surface defined by a single coordinate for which the scalar wave equation is separable. Elliptic cylinder coordinates (u, v, z) are related to cartesian coordinates (x, y, z) by the transformation

$$x = c_0 \cosh u \cos v, \quad y = c_0 \sinh u \sin v;$$

the metric coefficients are given by $h_u = h_v = c_0(\cosh^2 u - \cos^2 v)^{\frac{1}{2}}$. The coordinate system in Fig. 9 shows a conducting elliptic cylinder defined by $u = u_0$ with a slot of angular width $2v_1$, located at $v = v_0$. The coordinates u and v are radial and angular coordinates corresponding to r and φ of the circular cylinder.

The wave functions resulting from separation of the wave equation consist of products of radial and angular Mathieu functions. Using the series expansion of a plane wave in terms of even and odd elliptic wave functions as given by Stratton,[12] the transverse magnetic potentials inside and outside the cylinder $u = u_0$ can be written as follows:

$$\Pi_e^I = (8\pi)^{\frac{1}{2}} K \sum_{n=0}^{N} [(a_n^e/N_n^e) Je_n(c, \cosh u) \times Se_n(c, \cos v) Se_n(c, \cos\beta) + (a_n^o/N_n^o) Jo_n(c, \cosh u) \times So_n(c, \cos v) So_n(c, \cos\beta)]. \tag{25a}$$

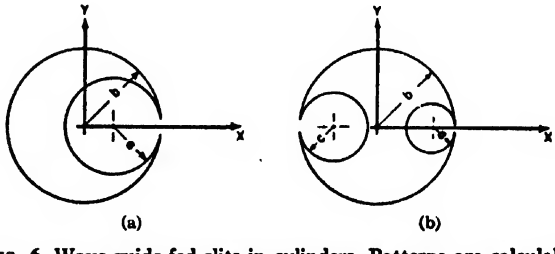


FIG. 6. Wave-guide-fed slits in cylinders. Patterns are calculable from slitted cylinder theory.

[12] Reference 4, p. 386, Eq. (84).

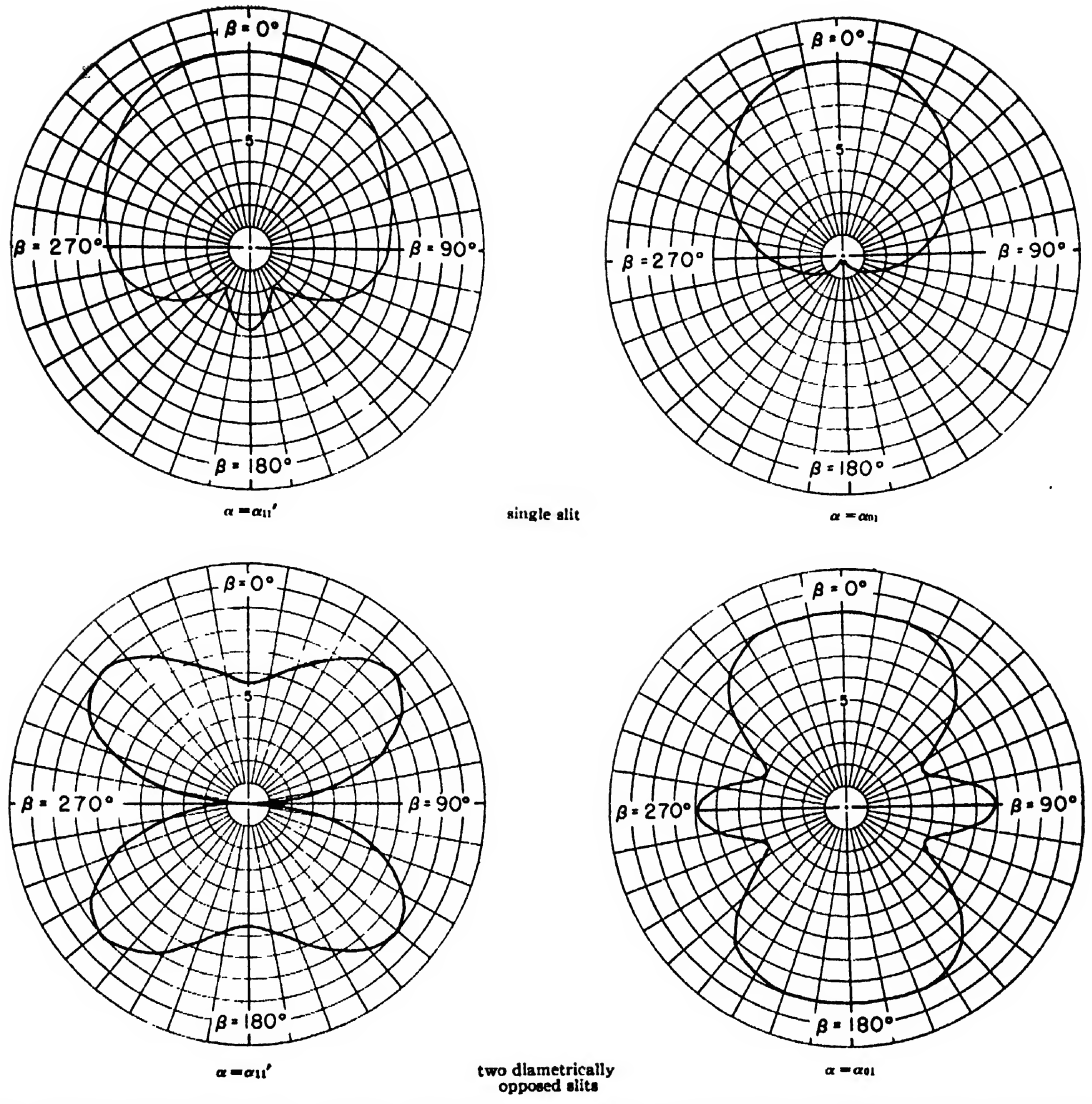


FIG. 7. Relative conical patterns of a circular cylinder containing one or two diametrically opposed slits when the cylinder is excited in the TE_{11} or TM_{01} circular wave-guide mode.

$$\Pi_z^{II} = (8\pi)^{\frac{1}{2}} K \sum_{n=0}^{N} \Big[\frac{1}{N_n^e} \{ i^n Je_n(c, \cosh u) + c_n^e He_n^{(2)}(c, \cosh u) \} \times Se_n(c, \cos v) Se_n(c, \cos\beta) + (1/N_n^o) \{ i^n Jo_n(c, \cosh u) + c_n^o Ho_n^{(2)}(c, \cosh u) \} \times So_n(c, \cos v) So_n(c, \cos\beta) \Big]. \quad (25b)$$

The boundary conditions require that Π_z match over the interval $0 \leq v \leq 2\pi$ and vanish over C_1, and that $\partial\Pi_z/\partial u$ match over C_2. The condition that Π_z must match over the interval of orthogonality of the angular functions requires that

$$a_n^e Je_n = i^n Je_n + c_n^e He_n^{(2)}, \quad (26a)$$

$$a_n^o Jo_n = i^n Jo_n + c_n^o Ho_n^{(2)}, \quad (26b)$$

where the radial functions are evaluated at $u = u_0$. The other two boundary conditions relate the coefficients over the non-orthogonal subintervals C_1 and C_2. Minimizing the total mean square error and letting v_1 vanish again results in a set of independent equations for the unknown coefficients a_n. When

$$Je_{a_e}(c, \cosh u_0) = 0, \quad (27a)$$

or

$$Jo_{a_o}(c, \cosh u_0) = 0, \quad (27b)$$

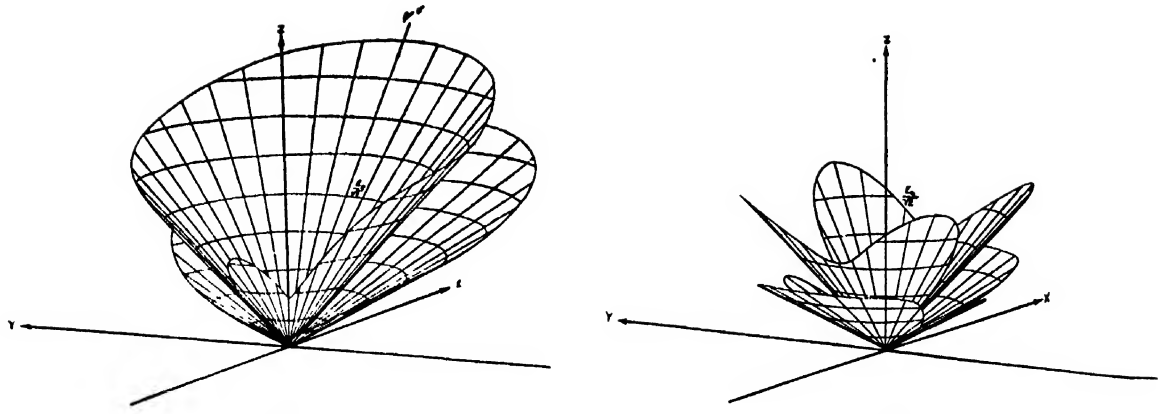


FIG. 8. Three-dimensional pattern of a radial electric probe located at $\varphi=(\pi/2)$ in a $\frac{1}{2}\lambda$ diameter cylinder containing (a) a single slit located at $\varphi=0$, and (b) two diametrically opposed slits located at $\varphi=0$ and $\varphi=\pi$. The incident wave is polarized such that $E_{0\alpha}=E_{0\beta}=(\sqrt{2}/2)E_0$. The patterns in the critical cones are taken from Fig. 7. Patterns at $\alpha=\alpha_{01}$ are shown multiplied by a factor of two.

the coefficients for $p\neq q$ are given by

$$a_p^{e}Je_p^{2}Se_p^{\beta}=\frac{2v_1Se_p^{v_0}}{He_p^{(2)}}F_e(c, v_0, \beta), \quad p\leq N, \quad (28a)$$

$$a_p^{o}Jo_p^{2}So_p^{\beta}=\frac{2v_1So_p^{v_0}}{Ho_p^{(2)}}F_e(c, v_0, \beta), \quad p\leq N, \quad (28b)$$

while that for $p=q_e$ is given by

$$a_{q_e}^{e}\frac{Se_{q_e}^{\beta}}{N_{q_e}^{e}}=-i\frac{Ne_{q_e}}{Se_{q_e}^{v_0}}F_e(c, v_0, \beta), \quad q_e\leq N, \quad (29a)$$

and that for $p=q_o$ is given by

$$a_{q_o}^{o}\frac{So_{q_o}^{\beta}}{N_{q_o}^{o}}=-i\frac{No_{q_o}}{So_{q_o}^{v_0}}F_e(c, v_0, \beta), \quad q_o\leq N, \quad (29b)$$

where $Se_{q_e}^{\beta}$ represents $Se_{q_e}(c, \cos\beta)$, $Se_{q_e}^{v_0}$ represents $Se_{q_e}(c, \cos v_0)$, etc., and where the elliptic conical pattern factor is given by

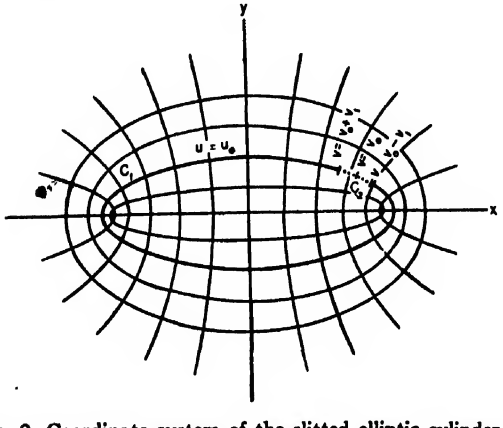


FIG. 9. Coordinate system of the slitted elliptic cylinder. The cylinder is defined by $u=u_0$ with distance between foci $2c_0$. A single slit is located at $v=v_0$ with angular width $2v_1$.

$$F_e(c, v_0, \beta)=\sum_{n=0}^{N} i^n\left(\frac{Se_n^{\beta}Se_n^{v_0}}{N_n^{e}He_n^{(2)}}+\frac{So_n^{\beta}So_n^{v_0}}{N_n^{o}Ho_n^{(2)}}\right). \quad (30)$$

The coefficients a_n^{e*} and a_n^{o*} for TE modes are obtained from those for TM modes by replacing the radial functions by their derivatives in Eqs. (28) to (30).

The fields in the cylinder at the critical angles ${}_e\alpha_{qm}$ and ${}_o\alpha_{qm}$ are the ordinary wave-guide modes in the elliptic wave guide.[13] At all angles other than the critical angles the fields vanish according to Eq. (28). The elliptic conical pattern factor in a plane normal to the axis is the same as that derived by Sinclair[14] and Carter[15] for the two-dimensional TE case. Since the incident angle α appears in the pattern factor only in the product $c=kc_0\sin\alpha$, the pattern in any cone $\alpha=\alpha_0$ about a cylinder $u=u_0$ whose distance between foci is $2c_0$ is the same as the pattern in a plane normal to the axis of a cylinder $u=u_0$ whose distance between foci is $2c_0\sin\alpha_0$.

The transition to the circular case follows as the limit when $c_0\to 0$ and $u\to\infty$, for then $c_0\cosh u\to r$. The solution for the elliptic cylinder then goes over to that for the circular cylinder.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to George Sinclair for his introduction to the theory of diffraction problems, and to V. H. Rumsey for his interest and constructive criticisms.

Work described in this paper was carried out, in part, under a contract between the Wright-Patterson Air Force Base of the Air Materiel Command and The Ohio State University Research Foundation.

[13] L. J. Chu, J. App. Phys. 9, 583–91 (1938).

[14] George Sinclair, "The Patterns of Antennas Mounted near Elliptic Cylinders," The Ohio State University Research Foundation Report No. 301–1, June, 1947.

[15] Carter, Martin, and Thorne, "Slot Antennas in Elliptic Cylinders and Airplane Vertical Stabilizer," RCA Laboratories Report No. CM-45-15, April, 1948.

Frequency Factor and Activation Energy for the Volume Diffusion of Metals*

G. J. DIENES
Atomic Energy Research Department, North American Aviation, Inc., Los Angeles, California
(Received May 12, 1950)

The theoretical equations proposed for calculating D_0 of the diffusion-constant-temperature relation

$$D=D_0e^{-E/RT}$$

are critically discussed in the light of presently available data. It is concluded that none of the relations gives satisfactory agreement with experiment.

Empirical analysis of experimental data covering a wide range of D_0 and E values showed that the quantity E/T_m, where T_m is the melting point in °K, is the main factor in determining the value of D_0, and that D_0 depends approximately exponentially on E/T_m. In addition to many other diffusion constants, it is shown that the anisotropic diffusion of zinc and the widely differing diffusion constants for self-diffusion in α- and γ-iron are included in the proposed empirical correlation.

A possible theoretical basis for the empirical correlation is described based on the idea of local melting or disordering. The quantity E/T_m is identified with the entropy of activation.

I. INTRODUCTION

IT is well known that the temperature dependence of diffusion can be well described by the simple exponential relation

$$D=D_0e^{-E/RT}, \tag{1}$$

where D=diffusion constant; D_0=constant, same dimensions as D; E=activation energy; R=gas constant; T=temperature in °K. Many attempts have been made to derive theoretical or semi-empirical expression for D_0, the so-called frequency factor. Recently, sufficient data on volume diffusion have become available to judge critically the validity of these equations.[1]

It may be pointed out that investigations over the last ten years or so have shown very clearly that the diffusion constant as well as the activation energy is a function of the concentration in most intermetallic systems. The importance of proper thermodynamic treatment has been emphasized (use of activities in place of concentrations). A review by A. D. Le Claire[2] summarizes well our present state of knowledge, and it is clear that the corresponding D_0 of such systems is not yet amenable to theoretical analysis. Any relation proposed for D_0, therefore, should be tested by comparison with self-diffusion data and with intermetallic diffusion data obtained at very low concentrations and far below the solubility limits so that concentration dependence is negligible. Since grain boundary and surface diffusion present special problems, the present discussion will be limited to volume diffusion.

It should also be remarked that evaluation of D_0 should not be tested by comparing experimental and calculated values of the activation energy, E. Such a calculation is found to be quite insensitive to the value of D_0 chosen because of the masking effect of the exponential factor in Eq. (1). D_0 values now known cover a sufficiently wide range for proper comparison of any proposed relation with experiment.

Table I is a summary of proposed theoretical and semi-empirical relations. (For notation, see headings of Table II.) In the third column the implied linear functional relations are also indicated. Leichter[3] has pointed out that only the Langmuir-Dushman[4] equation and the relation proposed by him predict the right order of magnitude for D_0. This conclusion is based on a few self-diffusion studies.[5] It is probably far more important, however, that none of the relations listed in Table I predict the correct functional dependence. Sufficient data are given in Table II for systems of known self-diffusion to verify this statement by simply plotting the data according to column three of Table I. Among others, glaring examples of deviations may be found in the anisotropy of Zn and in the self-diffusion of iron in the α- and γ-phase.[5]

Bradley's[6] relation comes closest to showing a correlation provided the relation between D_0 and Evd^2/T_M (where T_M is the mean temperature of the experiment) is taken to be an exponential rather than a linear one. It will become clear from later discussion in this paper that the above is to be expected for the simple reason that T_M is usually not too far removed from the melting point.

* This document is based on work performed under Contract No. AT-11-1-GEN-8 for the AEC.

[1] C. J. Smithells, *Metals Reference Book* (Interscience Publishers, Inc., New York, 1949), pp. 390–412.

[2] A. D. Le Claire, Section on Diffusion of Metals in *Progress in Metal Physics I* (Interscience Publishers, Inc., New York, 1949), pp. 306–379.

For surveys of the field of diffusion in metals see:
R. M. Barrer, *Diffusion in and Through Solids* (Cambridge University Press, 1941).
W. Jost, *Diffusion und Chemische Reaktion in Festen Stoffen* (Theodor Steinkopff Verlag., Dresden, 1937).
W. Seith, *Diffusion in Metallen* (Verlag. Julius Springer, Berlin, 1939).
F. Seitz, *The Physics of Metals* (McGraw-Hill Book Company, Inc., New York, 1943).

[3] M. Leichter, N.A.C.A. Tech. Note No. 1856, Washington, April, 1949.

[4] I. Langmuir and S. Dushman, Phys. Rev. 20, 113 (1922).

[5] C. F. Birchenall and R. F. Mehl, J. App. Phys. 19, 217 (1948).

[6] R. S. Bradley, Trans. Faraday Soc. 33, 1185 (1937).

TABLE I. Equations proposed for D_0 of $D=D_0e^{-E/RT}$.

Investigator	Proposed equation	Implied linear functional relation
Langmuir-Dushman[a]	$D_0=(E/Nh)d^2$	D_0 *vs.* Ed^2
Bradley[b]	$D_0=1/6(E/RT)\nu d^2$	D_0 *vs.* $E\nu d^2/T_m$
Van Liempt[c]	$D_0=8\nu d^2/3\pi$	D_0 *vs.* νd^2
Eyring[d]	$D_0=(V/N)^{\frac{1}{3}}(KT/h)(1-e^{-3\theta/4T})$	D_0 *vs.* $V^{\frac{1}{3}}\theta$*
Leichter[e]	$D_0/\nu d^2=2\pi W/3(3E_i/E_l)^{\frac{1}{2}}e^{E_f/12RT}$	$\mathrm{Log}D_0/\nu d^2$ *vs.* E_f/T_m**
Cichocki[f]	$D_0=2.43\times10^4(M)^{\frac{1}{2}}/V(T_m)^{\frac{1}{2}}$	D_0 *vs.* $(M)^{\frac{1}{2}}/V(T_m)^{\frac{1}{2}}$

* Exponential can be expanded for all values of interest here.
** Other quantities are essentially constant for values of interest here.
[a] See reference 4.
[b] See reference 6.
[c] J. Van Liempt, Zeits. f. Physik 96, 534 (1935).
[d] See reference 11.
[e] See reference 3.
[f] J. Cichocki, J. de phys. et rad. 7, 420 (1936).

II. EMPIRICAL CORRELATION

The discussion given so far indicates that no suitable expression exists (theoretical or empirical) for relating D_0 to, and for calculating it from, other physical quantities. An analysis of presently available experimental data was undertaken in order to find out whether it is possible to abstract from the data themselves the physically significant quantities which determine D_0. Guided by the often proposed idea of melting, or disorder, as an important factor in diffusion, it was empirically discovered that the physically important quantity is

$$E/T_m,$$

where E=activation energy; T_m=melting point in °K. This conclusion is based on the data given in Tables II and III. An attempt was made to include all reliable diffusion data within the restrictions discussed in the introduction. In cases where several different values are available, Smithell's judgment has been accepted in most cases as to reliability (as given in his tables).[1] It was found that good correlation is obtained over a wide range of D_0 values by plotting $\log(D_0/\nu d^2)$ or $\log(D_0/\nu\lambda^2)$ against E/T_m. The best correlation appears to be given by $\log(D_0/\nu\lambda^2)$ *vs.* E/T_m. The possibility of a close relation between E and T_m for self-diffusion has been discussed by many workers.[7] It is to be emphasized that the present discussion is concerned with the relation between D_0 and the above quantities, rather than the possible interrelation of E and T_m.

The results are shown in Fig. 1, where $\log(D_0/\nu\lambda^2)$ is plotted against $E/2.303RT_m$ for all the items of Table II and III. Several values are available for Cu with three typical ones given in Table II. All of these fall along the correlation curve and furnish no judgment as to the best values for this material. This is probably due to compensating errors in E and D_0 in the various experimental data. In Table III all physical constants refer to the solvent atoms. This is obviously incorrect in the immediate neighborhood of a solute atom but no data are available for more accurate analysis. These variations must be minor, however, relative to the exponential influence of E/T_m. Within the experimental reliability of D_0 values the correlation is seen to be a very good one and covers a D_0 range from about 10^{-9} to 10^{+4} and activation energies from 8 to 140 kcal./mole. This curve clearly establishes E/T_m as the main factor which determines the value of D_0. However, since the influence of this factor is approximately an exponential

TABLE II. Data for self-diffusion of metals.[a]

System	Frequency factor, D_0, cm² sec.⁻¹	$\log\frac{D_0}{\nu d^2}$	$\log\frac{D_0}{\nu\lambda^2}$	Activation energy, E, kcal./mole	Melting point, T_m, K°	$\frac{E}{2.303RT_m}$	Lattice constant, d, cm ×10⁻⁸	Interatomic distance, λ, cm ×10⁻⁸	Debye frequency, ν, sec.⁻¹ ×10¹²	Debye temp. θ	Av. T. of exp. T_m °K	Molar volume, V, cm³	Heat of fusion E_f kcal./mole
Silver	8.95×10^{-1}	2.08	2.38	45.9	1234	8.32	4.08	2.88	4.5	215	1110	10.8	2.73
Gold	1.6×10^{-1}	1.41	1.71	53.0	1336	8.68	4.07	2.87	3.8	170	1183	10.4	3.06
Copper ●	4.7×10^{1}	3.74	4.04	61.4	1357	9.89	3.61	2.55	6.6	315	1113	7.1	3.11
Copper[b]	1.1×10^{1}	3.11	3.41	57.0	1357	9.20	3.61	2.55	6.6	315	1113	7.1	3.11
Copper[c]	3.0×10^{-1}	1.54	1.84	46.8	1357	7.55	3.61	2.55	6.6	315	1113	7.1	3.11
Lead	6.6	3.17	3.48	27.9	600	10.15	4.94	3.48	1.8	88	488	18.3	1.19
Tungsten	1.15×10^{1}	3.24	3.3/	140.0	3655	8.37	3.16	2.73	6.6	310	~3000	9.55	11.2
Zinc, (∥) c-axis	4.6×10^{-3}	0.55	0.55	20.4	693	6.44	4.94	4.94	5.3	250	655	9.2	1.72
Zinc, (⊥) c-axis	9.1×10^{1}	4.39	4.39	31.0	693	9.77	2.65	2.65	5.3	250	655	9.2	1.72
Iron α[d]	3.4×10^{4}	6.65	6.74	77.2	1183[e]	14.23	2.86	2.58	9.4	420	1074	7.07	3.63
Iron γ[d]	1.0×10^{-3}	−1.08	−0.76	48.0	1808[f]	5.78	3.56	2.57	9.4	420	1297	7.07	3.63

[a] Data from reference 1 unless otherwise noted.
[b] From Steigman, Shockley, and Nix, Phys. Rev. 56, 13 (1939).
[c] From Raynor, Thomassen, and Rouse, Trans. Am. Soc. Met. 30, 313 (1942).
[d] Data from reference 5.
[e] This is the transition temperature from α to γ as explained in text.
[f] This is the melting point of iron. A somewhat lower melting point for γ-iron proper would not influence the results appreciably.

[7] See, for example, W. A. Johnson, Trans. A.I.M.E. 143, 107 (1941), also Table X of reference 2.

TABLE III. Data for intermetallic diffusion.*

System	Frequency factor, D_0, cm² sec.⁻¹	$\log\frac{D_0}{\nu d^2}$	$\log\frac{D_0}{\nu\lambda^2}$	Activation energy, E, kcal./mole	Melting point, T_m, °K	$\frac{E}{2.303RT_m}$	Lattice constant, d, cm ×10⁻⁸	Interatomic distance, λ, cm ×10⁻⁸	Debye frequency, ν, sec.⁻¹ ×10¹²	Debye temp. θ	Av. T. of Exp. T_{Av} °K
Cd in Ag	4.9×10^{-5}	−2.19	−1.89	22.4	1234	3.98	4.08	2.88	4.5	215	1045
Cu in Ag	6.0×10^{-5}	−2.10	−1.80	24.8	1234	4.39	4.08	2.88	4.5	215	1045
In in Ag	7.3×10^{-5}	−2.01	−1.72	24.4	1234	4.33	4.08	2.88	4.5	215	1045
Sb in Ag	5.3×10^{-5}	−2.15	−1.85	21.7	1234	3.85	4.08	2.88	4.5	215	1045
Sn in Ag	7.8×10^{-5}	−1.98	−1.69	21.4	1234	3.80	4.08	2.88	4.5	215	1045
Cu in Al	8.4×10^{-2}	0.79	1.10	32.6	933	7.65	4.04	2.86	8.34	398	784
Cu in Au	1.1×10^{-3}	−0.77	−0.47	27.4	1336	4.48	4.07	2.87	3.8	170	1000
Al in Cu	1.8×10^{-2}	0.31	0.61	37.7	1357	6.07	3.61	2.55	6.6	315	1073
Be in Cu	2.3×10^{-4}	−1.57	−1.27	28.0	1357	4.52	3.61	2.55	6.6	315	1073
Cd in Cu	2.0×10^{-9}	−6.62	−6.33	8.0	1357	1.29	3.61	2.55	6.6	315	1073
Si in Cu	3.7×10^{-2}	0.63	0.94	40.0	1357	6.44	3.61	2.55	6.6	315	1073
Sn in Cu	1.1	2.12	2.42	45.0	1357	7.25	3.61	2.55	6.6	315	1073
Cu in Ni	1.0×10^{-3}	−1.0	−0.70	35.5	1728	4.50	3.51	2.48	7.85	375	1043
Cd in Pb	1.8×10^{-3}	−0.39	−0.08	15.4	600	5.62	4.94	3.48	1.8	88	483
Sn in Pb	4.0	2.96	3.27	26.2	600	9.53	4.94	3.48	1.8	88	538
Ni in Pb	6.6×10^{-1}	2.18	2.49	25.3	600	9.21	4.94	3.48	1.8	88	575
Tl in Pb	2.5×10^{-2}	0.76	1.07	19.4	600	7.06	4.94	3.48	1.8	88	515
Ag in 50:50 AgAu**	3.9×10^{-1}	1.75	2.05	44.7	1300	7.52	4.07	2.87	4.1	190	1111
Au in 50:50 AgAu**	1.2×10^{-1}	1.24	1.54	44.1	1300	7.41	4.07	2.87	4.1	190	1111

* Data from reference 1 unless otherwise stated. Only data obtained at very low concentrations and well below the solubility limits included. Lattice parameters and other physical constants refer to the solvent.
** Data from reference 9.

one, more subtle differences, specifically the relation of D_0 to lattice parameters, are extremely difficult to determine until a large amount of considerably more accurate experimental information is available.

The correlation is seen to be non-linear, except perhaps over the middle range, where a line of slope 1, indicated by the broken line, may be drawn. Over the linear range D_0 may be expressed as

$$D_0 = \nu\lambda^2 K e^{E/RT_m} \qquad (2)$$

with $K = 10^{-6}$. Since the linear form is valid only over a limited range, no physical significance can be assigned to the value of K.

It is to be noted that the anisotropy of Zn is properly taken care of in this correlation.* Furthermore, self-diffusion in α- and γ-iron also fits into this scheme provided the transition point between the two phases is accepted as the melting point of α-iron. This choice is not unreasonable in view of the fact that such a transition involves a great loosening of the lattice over the transition range of temperatures. Bismuth, which also shows anisotropic diffusion,[8] has been left out of Table II and Fig. 1. Available data on Bi are very approximate.[8] In the direction perpendicular to the C axis, a tremendously high D_0 was obtained ($\sim10^{47}$), which can be fitted in approximately with the present correlation. In the parallel direction a low D_0 is obtained ($\sim10^{-3}$) and the corresponding activation energy appears to be much too high. This point would fit very poorly on Fig. 1; but, because of the large scatter in the

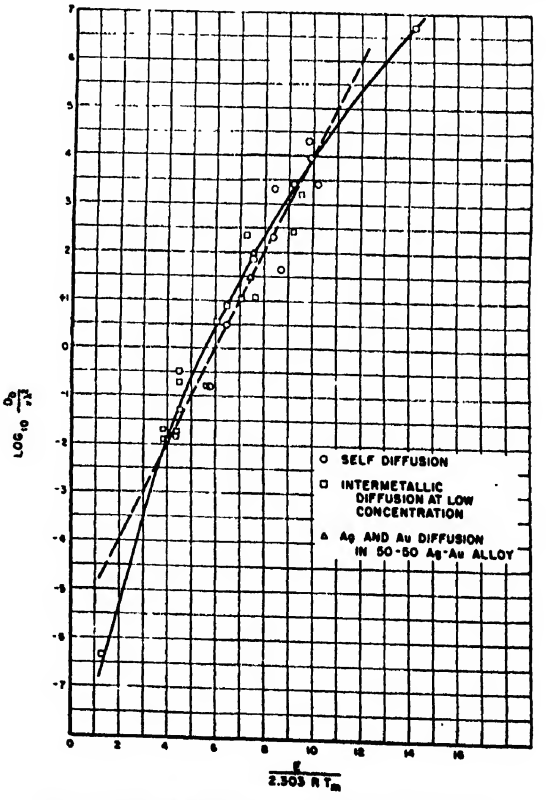


FIG. 1. Correlation of frequency factor with E/T_m.

* The Debye temperature given in Table II for Zn is an average value and, in the absence of better information, the same value was used for both directions. The characteristic temperature is unlikely to deviate by more than 30–40 percent from the average owing to anisotropy. This would have a negligible influence on the results since the frequency factors differ by more than a factor of 10.[3]

[8] W. Seith, Zeits. f. Elektrochemie 39, 538 (1933). See also R. F. Mehl, Trans. A.I.M.E. 122, 11 (1936).

data, it cannot be concluded that Bi is an exception to the correlation presented here. More experimental data on diffusion anisotropy as well as on the influence of phase changes (analogous to iron) would be very helpful.

It may be of interest also to note that Johnson's[9] values for the self-diffusion of silver and gold in a 50:50 AgAu alloy fall along the correlation curve.

III. THEORETICAL CONSIDERATIONS

It has been shown in the preceding sections that, empirically, the frequency factor, D_0, depends roughly exponentially on a definite fraction of the activation energy determined by the melting point. It is natural to look, therefore, for local melting, or disordering, as the elementary activation step. Such a proposal has been made by Mott[10] recently in connection with creep at low stresses. It appears that a similar picture may be applicable to diffusion. Let us assume that the elementary activation step is the disordering or melting of a small group of atoms by thermal fluctuations at a temperature T, which is below the melting point T_m, and let us assume that this small group of atoms may be treated thermodynamically. The free energy per atom necessary to produce this local melting, or disordering, may be written

$$\Delta F' = \Delta H' - T\Delta S' \tag{3}$$

with $\Delta F'$ equal to zero, at the melting point, T_m. Then

$$\Delta S' = \Delta H'/T_m. \tag{4}$$

The free energy of activation may be assumed to be proportional to $\Delta F'$ (the proportionality constant depending on the number of atoms involved),[10] i.e.,

$$F = \alpha\Delta F' = \alpha\Delta H' - T\alpha\Delta H'/T_m \tag{5}$$

or in the usual notation

$$F = E - TE/T_m, \tag{6}$$

where E/T_m represents the entropy of activation. Following reaction rate theory,[11] the expression for the diffusion constant D may be written

$$D = D_0 e^{-E/RT} = \beta e^{E/RT_m} e^{-E/RT} \tag{7}$$

which is of the same form as the empirical equation valid over the linear range of Fig. 1. D_0 is identified with

$$\beta e^{E/RT_m},$$

where $\beta = K\nu\lambda^2$.

Thus, it is possible to lend some theoretical support to the empirical analysis presented in this paper.

It may be added that Eq. (7) has no physical meaning for $T > T_m$. At $T = T_m$

$$D = D_m = K \cdot \nu\lambda^2. \tag{8}$$

The analysis presented here states, therefore, that melting temperatures represent equivalent states for comparison of diffusion processes. Physically, it is far more reasonable to compare diffusion constants, and to relate them to lattice parameters, at the melting point rather than at $T = \infty$ as implied by the use of D_0. It would be desirable, of course to know accurately the lattice parameters at the melting point.

Another feature of the picture presented here may be pointed out qualitatively. Experimental results up to date indicate[2] that the activation energy for the diffusion of any metal A in dilute solution in another metal B is always less than the activation energy for self-diffusion in B. Johnson[9] has pointed out that the hole mechanism of diffusion is apparently unable to account for this trend and proposed that diffusion takes place via the motion of substituted atom-hole complexes. This suggestion has been criticized by Frenkel[12] as being improbable.

The picture of diffusion presented here can explain the above phenomenon qualitatively. The free energy of activation in Eq. (5) may be lowered either by decreasing α, i.e., the size of the region to be melted, or by decreasing the heat of fusion, $\Delta H'$. It seems unlikely that α would be changed appreciably. The heat of fusion, however, may be lowered considerably even in a dilute solution of A in B, since it is precisely around an atom of A that local melting must occur, and the concentration of A in the small group of B atoms may be quite high. Some change in the melting point is also to be expected modifying somewhat at the correlation of Fig. 1. The wide applicability of this correlation shows, however, that the important change must be in the activation energy. No quantitative results can be given at this time.

With the importance of melting characteristics, as related to diffusion, empirically established, it seems likely that considerable further theoretical progress may be made by a careful investigation of melting and local disorder phenomena. Modern theories of melting and fluctuation phenomena should form a convenient starting point.

Note added in proof: Since this paper was written, the author's attention has been called to a recent paper by C. F. Birchenall and R. F. Mehl, Trans. A.I.M.E., 188, 144 (1950), in which new data are given for self-diffusion in α- and γ-iron which supersede entirely their previous results (Reference 5). These new data fit into the correlation scheme presented here and do not change the conclusions given in the text. As a matter of fact, since the activation energies for the two phases are now practically identical, the explanation in terms of melting points may acquire added significance.

[9] W. A. Johnson, Trans. A.I.M.E. 147, 331 (1942).
[10] N. F. Mott, Proc. Phys. Soc. London 60, 391 (1948).
[11] S. Glasstone, K. J. Laidler, and H. Eyring, *The Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941), pp. 519, 538.
[12] J. Frenkel, *Kinetic Theory of Liquids* (Oxford University Press, Oxford, 1946), pp. 27–36.

Letters to the Editor

Thorium Sulfide as a Thermionic Emitter

T. E. HANLEY
Naval Research Laboratory, Washington, D. C.
June 28, 1950

IN a recently declassified AEC report[1] it is stated that sulfides of thorium, particularly ThS, have high resistance to thermal shock and are "more resistant toward metals than the oxides." In addition, ThS has an electrical conductivity of $10^{-1}\Omega$ cm and "is as non-volatile as the most refractory oxides." In view of these characteristics it was of interest to investigate the possibility of using ThS as a thermionic emitter.

A sample of this material was obtained through the AEC. A coating bath was set up in order to obtain uniform coatings. The bath was prepared by ball-milling (Pyrex balls in Pyrex container) 5 g of ThS with a mixture of 50 ml of methanol and 50 ml of acetone. This combination when used as a coating bath with a nickel anode would put a 0.001-in. coating on a 0.0005-in. wire in 10 sec. at 50 v and an initial current density of 40 ma/cm^2.

The temperature scale for ThS was obtained by heating a ThS-coated 0.002-in. molybdenum cylinder (o.d. = 0.125 in., $L = 3.5$ in.) in vacuum. The brightness temperature was measured on the outer coated surface, while the true temperature was obtained by sighting the optical pyrometer on a 0.30-in. hole in the cylinder. The measurements give a spectral emissivity of 0.4 for ThS.

The thermionic emission data were taken using coated 0.005-in. tungsten wires in a guard ring diode structure. The Richardson constants of the material are approximately $\Phi = 3.4$ ev and $A = 100$ amp./cm^2/deg.^{2}. In the vicinity of 1500°C Br the emission from ThS is only one-fourth as great as that from thoria (ThO_2).

Some measurements were also made to determine the rate of evaporation of this material by measuring the loss of weight from a uniformly coated tungsten strip which was held at a known temperature for a given length of time in vacuum. These measurements indicate a vapor pressure of 10^{-3} mm Hg at 1870°C Br whereas the previously mentioned report[1] gives the vapor pressure as less than 10^{-3} mm Hg at 2200°C. Perhaps comparison of these data is not in order, since the conditions of the measurements for the values given in the AEC report are not stated.

[1] Brewer, Bromley, Giles, and Lofgren, "The preparation and properties of refractory sulfides," University of California (November, 1945).

On the Fredrickson-Eyring Theory of the Mechanical Behavior of Metals*

J. FLEERMAN† AND G. J. DIENES
Atomic Energy Research Department, North American Aviation, Inc., Downey, California
August 28, 1950

IT is generally felt that a unified theory of the mechanical behavior of metals must be based upon a dynamical theory since it is well known that the results of a stress-strain experiment depend markedly upon the time scale of the experiment. The first serious attempt to formulate a dynamical theory of the behavior of metals is due to Fredrickson and Eyring.[1] We shall refer to this paper as FE. It is the object of this letter to discuss some of the implications of their theory.

FE have used the method of the statistical rate theory of processes to formulate a theory of the mechanical behavior of metals. The behavior of the fundamental unit of dashpot and spring is determined by the differential equation

$$d\epsilon/dt = (1/g_i)(d\sigma/dt) + k_i \sinh\alpha_i\sigma, \qquad (1)$$

where σ and ϵ are the stress and strain, respectively, g_i is the spring constant of the ith element and, k_i and α_i are the viscosity parameters determined by way of a quantity β_i. The first term on the right-hand side of the equation describes the elastic properties of metals according to Hooke's law; the last term describes the flow properties according to the Eyring theory of viscosity. As a model used to interpret a stress-strain curve, FE use the combination of springs g_1, g_2, and g_3 and dashpot elements β_1 and β_2 in the form of Maxwell elements coupled in parallel. The spring g_1 and dashpot β_1 combine to form unit one, unit two consists of the spring g_2 and dashpot β_2, unit three is the spring element g_3 alone. FE have applied the results of their analysis to determine the equation for a loading curve obtained with a constant rate of straining.

To test the generality of this dynamical theory we have used Eq. (1) to calculate an unloading curve and (what turns out to be more interesting) a reloading curve calculated from the assumption that the material is reloaded immediately after unloading. The problem of strain relaxation, or recovery, at zero load is discussed separately under several simplifying assumptions.

With the test material loaded to a stress $\sigma_0^{(1)}$ at a strain $\epsilon_0^{(1)}$ under the conditions of a constant strain rate $\dot{\epsilon}$ the unloading curve can be evaluated by solving Eq. (1) subject to the initial condition $\sigma^{(1)} = \sigma_0^{(1)}$ at $t = 0$. We get

$$\sigma^{(1)} = \sigma_0^{(1)} - \frac{1}{\alpha}\ln\left\{\frac{1+\left[1+\dfrac{\exp(\alpha\sigma_0^{(1)})}{\beta}\right]\tanh\left[\dfrac{\alpha g}{2}(\epsilon_0^{(1)}-\epsilon)\right]}{1-\tanh(\alpha g/2)(\epsilon_0^{(1)}-\epsilon)}\right\}. \qquad (2)$$

The unloading curve for each unit of the Maxwell representation is calculated from Eq. (2) to a value of zero stress for the entire model. It turns out that after unloading the first unit is in a state of compression while the second and third units are in extension.

Similarly, the reloading curve can be determined from Eq. (1) using the initial condition that at $t = 0$, $\sigma = \sigma_0^{(2)}$, where $\sigma_0^{(2)}$ can be positive or negative depending upon whether it refers to unit one or two. We get

$$\sigma^{(2)} = \sigma_0^{(2)} + \frac{1}{\alpha}\ln\left\{\frac{1+\left[1+\dfrac{\exp(-\alpha\sigma_0^{(2)})}{\beta}\right]\tanh\left[\dfrac{\alpha g}{2}(\epsilon-\epsilon_0^{(2)})\right]}{1-\left[1-\dfrac{\exp(\alpha\sigma_0^{(2)})}{\beta}\right]\tanh\left[\dfrac{\alpha g}{2}(\epsilon-\epsilon_0^{(2)})\right]}\right\}. \qquad (3)$$

We have calculated the unloading and reloading stress-strain curves for SAE 1020 steel according to Eqs. (2) and (3), respectively,

1. at $T = -70$°C and a strain rate of 5×10^{-5} sec.$^{-1}$,
2. at $T = -70$°C and a strain rate of 500×10^{-5} sec.$^{-1}$,
3. at $T = 100$°C and a strain rate of 5×10^{-5} sec.$^{-1}$,

where we have assumed the material has been loaded to a value of true strain equal to 0.5 and then unloaded.

The results of these calculations are shown in Fig. 1. It will be noted that in all cases the reloading curve rises appreciably above the initial loading curve for equivalent values of strain. It appears that the model predicts too much strain hardening.[2]

It is important to determine the magnitude and rate of strain recovery to be expected from the FE theory. Such a calculation is to be carried out via Voigt-type of mechanical model representation. However, as a result of the non-linear character of Eq. (1), the equations which arise in the transformation of a Maxwell representation to a Voigt representation are too complex to interpret in a simple manner. In order to make an estimate of the relaxation time we have replaced the non-linear viscous element in the Maxwell representation by an equivalent Newtonian viscous element η. If we choose the viscosity η to be given by the equation

$$\eta = 1/\alpha k, \qquad (4)$$

then the relaxation time which is calculated will always be a time that is greater than any relaxation time obtained from the FE theory. Furthermore, it can be shown that another simplification which may be made, without appreciably influencing the results is the following: letting $g_1 = 0$ in the Maxwell representation renders the equivalent Voigt representation, consisting of springs

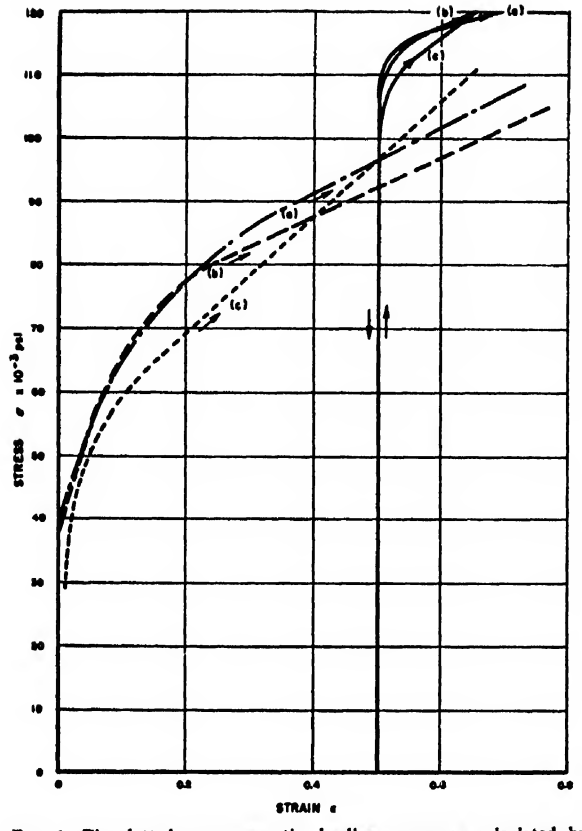


FIG. 1. The dotted curves are the loading curves as calculated by Fredrickson and Eyring. Curve (a) computed for $T=-70°C$ and $\dot{\epsilon}=500\times10^{-6}$ sec.$^{-1}$, curve (b) computed for $T=-70°C$ and $\dot{\epsilon}=5\times10^{-6}$ sec.$^{-1}$, curve (c) computed for $T=100°C$ and $\dot{\epsilon}=5\times10^{-6}$ sec.$^{-1}$. The solid curves are the unloading and reloading curves computed from a strain $\epsilon=0.5$ as discussed in the text.

g_1' and g_2' and Newtonian viscous elements η_1' and η_2' easily manageable.

Using the method outlined, we have calculated the maximum relaxation time for SAE 1020 steel at $T=-70°C$ and $\dot{\epsilon}=5\times10^{-5}$ sec.$^{-1}$ to be 2.6 days and 2.2 hours for the two units which appear in the Voigt representation. It is evident that strain relaxation is going to take place at a rate far greater than experimentally expected for steel in a temperature range where plastic deformation should be essentially permanent.[3]

Thus, upon the basis of calculations presented here, we conclude that the FE dynamical theory of the mechanical behavior of metals has inherent within it the following characteristics:

1. It predicts too much strain hardening.
2. It predicts a relaxation time for deformation recovery which is much too small and consequently it fails to account for the permanent deformation in steel in the temperature range $-70°C$ to $100°C$.

* This letter is based on studies conducted for the AEC under Contract AT-11-1-GEN-8.
† Now at the National Bureau of Standards, Washington, D. C.
[1] J. W. Fredrickson and H. Eyring, *Statistical Rate Theory of Metals*, Metals Tech. TP 2423, August, 1948.
[2] See for example: F. Seitz, *The Physics of Metals* (McGraw-Hill Book Company, Inc. New York, 1943), Fig. 46, p. 72; Fig. 51, p. 83.
[3] A. M. Freudenthal and M. Reiner, J. App. Mech. 15, 265 (1948).

Base-Metal Effects in Thoria-Coated Filaments

HERBERT NELSON
Tube Department, Radio Corporation of America, Harrison, New Jersey
August 11, 1950

THE thoria-coated filament has shown promise of meeting the demands for a thermionic source supplying long-life emission at high current densities and voltages better than does either the barium-strontium oxide-coated cathode or the thoriated tungsten filament. Investigation of the properties of this new type filament has therefore appeared fruitful and has been carried out in various laboratories of the RCA organization. The purpose of the present communication is to summarize briefly some of the results obtained in the part of this investigation which has had to do with base-metal effects. Adherence of the thoria to the metal base, ease of thermionic activation, and level of thermionic emission reached have been studied as a function of base material.

With regard to adherence, a problem is created as a result of the rather large difference in thermal expansion between thoria and the refractory metals suitable for base material. The coefficient of thermal expansion for thoria[1] is 11.3×10^{-6}, while for W, Mo, Ta, and Pt, the values are 4.2, 4.9, 7.0, and 9.0×10^{-6}, respectively. On the basis of these figures it is to be expected that adherence is best with Pt and poorest with W. Experimental work has shown the foregoing to be true. Adherence was found poor with W, fair with Mo, good with Ta, and excellent with Pt. With Ta, layers of thoria as thick as 3 mils (45 mg/cm^2) were found to adhere to the end of 3000-to-4000-hour life tests conducted in diodes. With W, on the other hand, even very thin layers (8 mg/cm^2) did tend to chip off at the very beginning of life tests. With Mo, chipping difficulties were also encountered though filaments with as much as 20 mg/cm^2 of thoria were operated successfully on life for as long as 2000 to 3000 hours.

Though, as suggested above, the better match in thermal expansion may account for the superior adherence to Pt and Ta, it is also possible that the relatively high ductility of these metals at operating temperatures may be a contributing factor.

Since it seems reasonable to expect that for low work function, thoria, like BaSrO, depends upon a stoichiometric excess of metal in the crystal lattice, tests were made to determine whether the base metal, as a reducing agent, has an effect upon the ease of activation of thoria-coated filaments. For this purpose four tubes were constructed identical in all respects except that two were made with a base material of Ta while the other two were made with Pt. The tubes were exhausted and processed in identical fashion, the electron emission being determined as a function of the filament temperature during the initial glowing. The results obtained are shown in Fig. 1, curves *A* and *B* expressing the temperature-electron emission relation for thoria on Ta and Pt, respectively. The fact that high thermionic activity is reached at

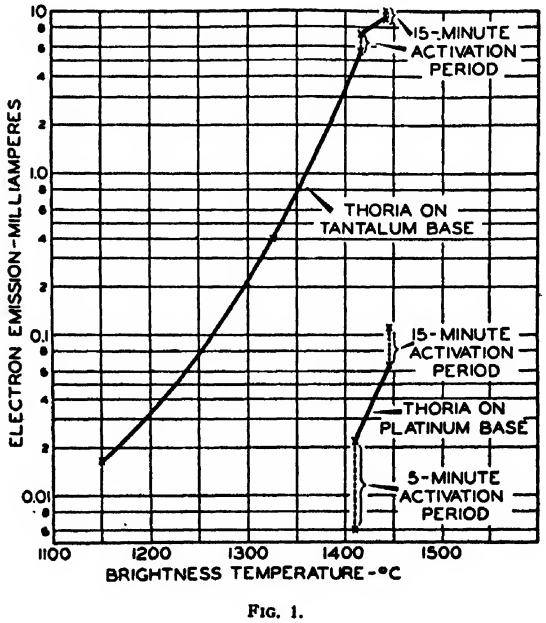


FIG. 1.

a greatly lower temperature with Ta, a good reducing agent, than with Pt, a poor reducing agent, is consistent with the view that a low work function of thoria is associated with a stoichiometric excess of metal in the crystal lattice.

When Mo, W, or Ta is used as base material, roughly the same maximum space-charge-limited emission, about 5 amp./cm², is obtained at 1900°K B.T. At lower temperatures, however, higher values are invariably obtained with Mo than with Ta. At 1800°K B.T., for instance, the average emission with Mo was found to be equal to 3.1 amp./cm², while the average with Ta was found to be 2.0 amp./cm². No explanation has as yet been found for this superiority of Mo at low temperatures.

Acknowledgment is made to the following members of the RCA organization who have made fundamental contributions to the development of the thoria-coated filament: Dr. Lloyd Garner, Dr. Philip Smith, and Mr. M. N. Fredenburgh

[1] A. Eisenstein, J. App. Phys. 17, 443 (1946).

Erratum: The Mechanical Properties of Glass

[J. App. Phys. 13, 623–634 (1942)]

F. W. Preston
The Preston Laboratories, Butler, Pennsylvania

On p. 631 there is an error in the caption for Fig. 5. The limiting velocity of the propagation of cracks should be 5080 ft./sec., rather than 3 miles/sec. as stated.

A similar error occurs on p. 624, in which the velocity of sound in glass should be approximately 1 mile/sec., instead of 3 miles/sec.

Scheme for Analyzing Capillary Measurements on Non-Newtonian Liquids

Orrin H. Clark and M. L. Deutsch
Socony-Vacuum Laboratories, Research and Development Department, Paulsboro, New Jersey

August 11, 1950

It has kindly been pointed out by Mr. A. B. Bestul that the formulas for analyzing non-Newtonian capillary flow given in our recent paper, "Scheme for analyzing capillary measurements on non-Newtonian liquids,"[1] were developed by B. Rabinowitsch, "Über die Viskosität und Elastizität von Solen,"[2] and have since been referred to in various other journals. The absence of these useful equations from such a book as Green's *Industrial Rheology and Rheological Structures* is surprising and led to the regrettable submission by us of their derivation as original material.

[1] Orrin H. Clark and M. L. Deutsch, J. App. Phys. 21, 713 (1950).
[2] B. Rabinowitsch, Zeits. f. physik. Chemie A145, 1–26 (1929).

Original Authorship of Formulas Given in Clark and Deutsch's Paper

M. Reiner
Institute of Technology, Haifa, Israel
and
G. W. Scott Blair
National Institute for Research in Dairying, University of Reading, England

September 14, 1950

The fact that the paper by Clark and Deutsch in the July issue of your Journal (p. 713) was passed by your referee as describing "a powerful new scheme" shows that rheological literature is not as widely known as it should be. May we, therefore, as authors of textbooks on rheology in which the relevant papers are quoted (Scott Blair, 1938; Reiner, 1949) be permitted to point out that this method is due to Weissenberg and was first applied by B. Rabinowitsch (1929) for the case of the capillary viscometer, followed by Mooney (1931), who adapted it to the rotating cylinder instrument? Hersey (1932) named it the "differentiation method" because it applies differentiation to the observed consistency curve in order to find the rheological equation of the material, in contradistinction to the "integration method," where a rheological equation is assumed and the consistency curve derived by integration.

We do not doubt that Messrs. Clark and Deutsch derived their Eq. (12) independently, but it will be seen that it is identical with Eq. (9) of Hersey's paper.

Application of Liouville's Approximation to the Blind Navigation Problem

J. J. Gilvarry and S. H. Browne*
The Rand Corporation, Inc., Santa Monica, California

September 8, 1950

In a recent paper[1] in this journal, it has been shown that the problem of determining a vehicle's position vector $\mathbf{r}$ from the reading $\mathbf{b}(t)$ of an accelerometer as a function of time t is equivalent to solving the differential equation

$$(d^2\mathbf{r}/dt^2)+\omega_0^2(R_0/R)^3(R_0\mathbf{k}+\mathbf{r})=\mathbf{b}(t), \tag{1}$$

where R (with initial value R_0) is the vehicle's radial distance from the earth's center, $\mathbf{k}$ is a radial unit vector at the origin, and $\omega_0=(g_0/R_0)^{\frac{1}{2}}$ if g_0 is the gravitational acceleration at the initial point of motion. A case not treated in that discussion presents itself when the vehicle's altitude $h=R-R_0$ relative to the earth is available as a function $h(t)$ of time from continuous measurements made by an altimeter (of radar or atmospheric-pressure type). The quantity $h(t)$ is then an independently measured integral of the motion (thus the problem of determining position coordinates is actually overdetermined). In this case, Eq. (1) becomes

$$(d^2\mathbf{r}/dt^2)+\omega_0^2\Gamma(t)\mathbf{r}=\mathbf{b}(t)-g_0\Gamma(t)\mathbf{k}, \tag{2}$$

where

$$\Gamma(t)=[1+h(t)/R_0]^{-2}\sim 1-3h(t)/R_0, \tag{3}$$

and the approximation indicated in (3) is valid for $h\ll R_0$. Equation (2) is linear in the coordinates and, from the standpoint of computer design, presents definite advantages over the nonlinear Eq. (1).

Any numerical method of solving a linear differential equation of second order will serve as a computational method to determine $\mathbf{r}(t)$ from a given $\mathbf{b}(t)$ and $h(t)$, or, when suitably mechanized, will serve as the basis of a computer design. For numerical computation, however, the majority of such methods are not well adapted to equations of the type (2), where one term (the $\mathbf{g}$-term) is a slowly varying function, but another term (the $\mathbf{b}$-term) may show rapid variations in time. When applicable, though, Liouville's approximation[2] is well suited to problems of this kind, and this letter indicates briefly its application to (2). The approximation in question is known as the Wentzel-Kramers-Brillouin-Jeffries approximation in its application as an asymptotic approximation in quantum mechanics; more generally, it appears in connection with transformations and asymptotic developments of Sturm-Liouville systems.[3] It is ascribed to Liouville by Schelkunoff.[4]

The quantity Γ (which is always positive for $|h|<R_0$) is approximately unity except for the perturbing parameter $h(t)$. Under the transformation

$$t'=\int_0^t \Gamma^{1/2}(\sigma)d\sigma, \quad \mathbf{r}'=\mathbf{r}\Gamma^{1/4}(t), \tag{4}$$

Eq. (2) yields the approximation

$$(d^2\mathbf{r}'/dt'^2)+\omega_0^2\mathbf{r}'=\Gamma^{-3/4}(t)\mathbf{b}(t)-g_0\Gamma^{1/4}(t)\mathbf{k}, \tag{5}$$

provided the first and second derivatives of $h(t)$ are sufficiently small so that

$$\Gamma^{-3/4}(d^2\Gamma^{-1/4}/dt^2)\ll\omega_0^2, \tag{6}$$

and hence the term $\Gamma^{-3/4}(d^2\Gamma^{-1/4}/dt^2)\mathbf{r}'$ can be neglected in (5). When the right-hand side of Eq. (5) is considered a function $\mathbf{b}'$

of t', the equation is linear with constant coefficients in the reduced variables t', r'. The explicit solution for r then follows directly as

$$r=(v_0/\omega_0\Gamma^{1/4})\sin\left[\omega_0\int_0^t\Gamma^{1/2}(\sigma)d\sigma\right]+\frac{1}{\omega_0\Gamma^{1/4}}$$
$$\times\int_0^t\left[\frac{b(\tau)}{\Gamma^{1/4}(\tau)}-g_0\Gamma^{3/4}(\tau)k\right]\sin\left[\omega_0\left(\int_0^t\Gamma^{1/2}(\sigma)d\sigma-\tau\right)\right]d\tau, \quad (7)$$

under the initial conditions $r_0=0$ and $(dr/dt)_0=v_0$. In actual application, Γ in (7) can usually be replaced by the approximation indicated in (3).

Equation (7) yields a convenient solution for numerical computation, since the solution is reduced to quadratures which can be carried out by the trapezoid or Simpson's rule (the sine function under the integral sign must be expanded by the addition theorem). To apply the method over an extended interval (over which the effect of the neglected term in Eq. (5) may be significant), one can use (7) as the step solution in the method of "linear continuation" described in the previous paper.[1] Alternatively, Schelkunoff's "wave perturbation" method[4,5] (with a slight modification for the presence of a forcing function) may be used; this method is based on Liouville's approximation, but defines the final solution as the limit of an iterative process over the over-all interval.

* Now at North American Aviation, Inc., Los Angeles, California.
[1] Gilvarry, Browne, and Williams, J. App. Phys. 21, 753 (1950).
[2] J. Liouville, J. de Math. 1, 253 (1836); 2, 16 (1837); 2, 418 (1837).
[3] R. Courant and D. Hilbert, *Methoden der mathematischen Physik* (Verlag. Julius Springer, Berlin, 1931), Vol. I, pp. 250, 290.
[4] S. A. Schelkunoff, Q. App. Math. 3, 348 (1945).
[5] M. C. Gray and S. A. Schelkunoff, Bell Syst. Tech. J. 27, 350 (1948).

Natural Crystal Greenockite (Cadmium Sulfide) as an X-Ray Detector

E. L. Criscuolo and D. T. O'Connor
U. S. Naval Ordnance Laboratory, Silver Spring, Maryland
September 8, 1950

RECENTLY,[1] artificially produced cadmium sulfide crystals have been described as sensitive x-ray detectors. Previous work described in the literature[2] refers to artificial cadmium sulfide crystals.

Natural greenockite crystal fragments of the order of 1 mm³ have been observed to pass direct currents up to 1×10^{-8} amp. in a 100-kv x-ray beam of an intensity of 4 roentgens per minute. This rough measurement was made using a micro-ammeter and a collecting battery voltage of 22 volts. Four samples gave similar results; they were light-sensitive and non-luminescent.

The mineral specimen was obtained from the Smithsonian Institute through the courtesy of Dr. W. F. Foshag, Chief of the Division of Mineralogy and Petrology, and Dr. R. Kellogg, Director.

[1] Rudolf Frerichs, J. App. Phys. 21, 312-317 (1950).
[2] Robert Hofstadter, Nucleonics (April, 1949); (May, 1949).

Measurements on the Diffusion of Interstitial Atoms in B.C.C. Lattices*

C. A. Wert
Institute for the Study of Metals, University of Chicago, Chicago, Illinois
August 14, 1950

THE alloys of interstitial dissolved atoms in the b.c.c. lattices offer the possibility that one can make extremely accurate measurements of the rate of diffusion of the solute atoms. This is so because the high temperature bulk diffusion measured by standard metallurgical methods can be extended to lower temperatures by measurements made on an atomistic scale by use of relaxation measurements. The diffusion of C in α-Fe offers an example of this technique. Stanley[1] measured the rate of diffusion in the range 500°C to 700°C. The author[2] measured diffusion in the range −35°C to +200°C. Together the data extend the knowledge of the diffusion coefficient over some 14 cycles of 10; this range is somewhat higher than has been done for any other system.

In an effort to extend this type of measurement to other systems, the author has made measurements of diffusion of C in Ta and Cb and of N in α-Fe. These measurements were made by two manifestations of the anelastic relaxation effects; the internal friction and the elastic aftereffect. The methods used are those described earlier.[2] The data are shown in Fig. 1. Here the diffusion

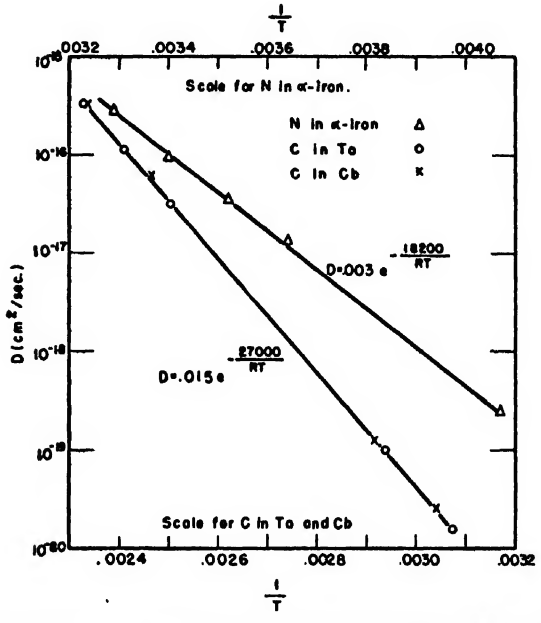


FIG. 1. Diffusion coefficient for C in Ta and Cb and N in α-Fe *vs.* $1/T$.

of N in α-iron and of C in Ta and Cb are seen to satisfy the well-known relation $D=D_0\exp(-H/RT)$. These data, of course, do not have the range quoted above for C in α-Fe; further measurements at much higher temperatures are needed for really accurate determination of D_0 and H.

An attempt to examine the various factors which make up D_0 has been made by Wert and Zener.[3] They expressed D_0 as

$$D_0=n\alpha a^2\nu e^{\Delta S/R}, \quad (1)$$

where n, α, and a^2 are geometrical constants for a given lattice, ν is the frequency of oscillation of a solute atom in an interstitial site, and ΔS an "entropy of activation." They found that if all of the activation energy for diffusion went into straining the lattice, then

$$\Delta S\approx -H(d\ln\mu/dT),$$

where μ is the shear modulus of the lattice. For simplicity one can rewrite this expression by defining $\beta=d(\mu/\mu_0)/d(T/T_m)$ obtaining

$$\Delta S\approx\beta H/T_m. \quad (2)$$

In this expression μ_0 is the modulus at $T=0$ and T_m is the melting temperature. β may be obtained for a large number of materials from some work of Köster.[4] For Ta and α-Fe the values of β are 0.40 and 0.43, respectively.

It follows from Eq. (1) that once D_0 is known from experiment, ΔS may be calculated. It is also true that if all the energy of activation goes into lattice strain energy, then ΔS is given by Eq. (2). A comparison of $\Delta S/R$ calculated by the two methods is given in Fig. 2. Here are shown data for C in α-Fe and Ta, N in

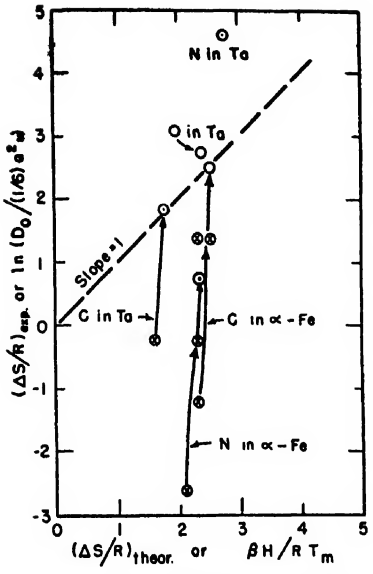


FIG. 2. $(\Delta S/R)$ as computed from Eq. (1) using diffusion experimental data *vs.* $\Delta S/R$ as computed from Eq. (2). For each material successive points are taken from the literature and the present letter as follows.

Oxygen in Ta	T. Ke, Phys. Rev. **74**, 9 (1948).
Nitrogen in Ta	T. Ke, Phys. Rev. **74**, 914 (1948).
Carbon in Ta	T. Ke, Phys. Rev. **74**, 9 (1948); C. A. Wert (present letter).
Carbon in α-Fe	J. Snoek, Physica **8**, 711 (1941); J. K. Stanley, Trans. A.I.M.E. **185**, 752 (1949); C. Wert and C. Zener, Phys. Rev. **76**, 1169 (1949); C. Wert, Phys. Rev. (to be published)
N in α-Fe	J. Snoek, Physica **8**, 711 (1941); C. Wert and C. Zener, Phys. Rev. **76**, 1169 (1949); C. Wert (present letter).

α-Fe and Ta, and O in Ta. The open circles represent the best data available at present; the other circles represent earlier data as given in the caption to Fig. 2. If $\Delta S/R$ computed by Eqs. (1) and (2) were identical, then all the points should lie on the dashed line. The trend of the measurements toward better agreement between $\Delta S/R$, as computed by the two methods, is quite apparent. This trend toward better agreement presumably results from the better values of D_0 obtainable by extension of the measurements over a wider range of temperature.

* This research was supported in part by the ONR (Contract No. N-6ori-20-IV, NR 019 302).

[1] J. K. Stanley, Trans. A.I.M.E. **185**, 752 (1949).
[2] C. Wert, Phys. Rev. (to be published).
[3] C. Wert and C. Zener, Phys. Rev. **76**, 1169 (1949).
[4] W. Köster, Zeits. f. Metallkunde **39**, 1 (1948).

A Remark about Standardization and "A Dimensional Analysis of Metal Cutting" *

ANDRÉ MARTINOT-LAGARDE
Associate Professor of Mechanics of Fluids, Lille University, France
September 8, 1950

ANALOGOUS non-dimensional variables appear in both metal cutting and mechanics of fluids. Therefore, it seems convenient to choose exactly the same non-dimensional variables. For instance in mechanics of fluids $c\rho Vl/K$ (c specific heat, ρ mass density, V velocity, l length, K thermal conductivity) is called the "Péclet number." This note is to propose writing the same number for metal cutting, instead of its inverse, used by Drucker and Ekstein, and using the same designation also. The velocity indeed is the variable which varies the most frequently, and it is convenient that every non-dimensional variable be proportional to such a quantity.

* D. C. Drucker and H. Ekstein, J. App. Phys. **21**, 104–107 (1950).

Tearline Patterns in Ferrochromium

C. A. ZAPFFE AND F. K. LANDGRAF
Baltimore, Maryland
August 14, 1950

IN the course of studying fractographic patterns,* special formations have frequently been observed within the fracture traverse which are believed to relate to conditions of original crystal growth. Some suggestively similar patterns, however, are discussed from another viewpoint in the recent paper by Kies, Sullivan, and Irwin in this journal.[1] Their contribution is an important advance in the interpretation of fracture patterns, and specifically those previously classified as Type II expressing transient conformations of the fracturing stress.[2] It also focuses attention upon a principal current problem in the interpretation of certain fracture patterns—the distinguishing between Types I and II. Many fractographs, without possible question, reveal structural features of the solid; but others clearly refer only to stress. Because the correct identification often has important implications for theories of the solid state, several complex examples are submitted here, taken from a study of chromium-rich iron.

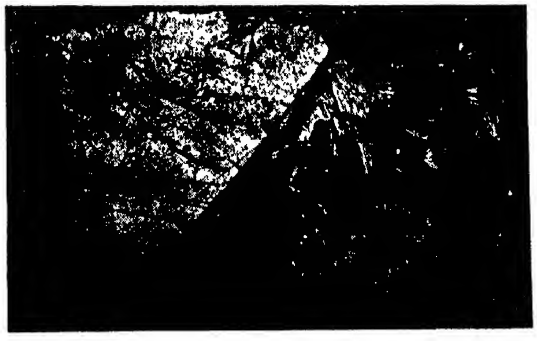


FIG. 1. Fractograph of ferrochromium containing 70 percent Cr and 0.05 percent C, showing a transition from transgranular to intergranular fracture traverse.

In Fig. 1, transgranular and intergranular fractures lie within the same general plane of focus. To the left of the diagonal boundary, fracture has followed a cleavage plane of the icositetrahedron {112}, displaying a relatively flat traverse with traces of intersecting planes which can be identified as also belonging to the icositetrahedral form. The traverse then shifts from a planar to a banal path along grain or lineage boundaries.

In Fig. 2 some of the fanlike forms visible on the banal traverse of the previous figure are shown in more detail. The condition here is particularly interesting because of an irregular boundary roughly bisecting the field. Patterns somewhat similar to these, as has been shown by Kies, Sullivan, and Irwin, can be obtained

FIG. 2. Intergranular surfacial detail in fractured 70 percent ferrochromium containing 0.17 percent C.

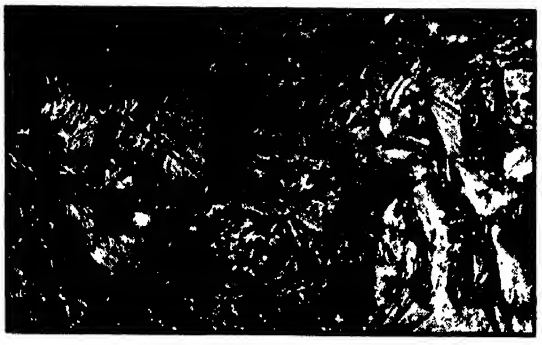


FIG. 3. Fractograph of ferrochromium containing 0.75 percent nitrogen, showing a boundary [lightly retouched for reproduction] which sharply separates the upper and lower halves of the field.

by special stress conformations during fracture of amorphous or rheologically isotropic bodies. However, solid metals show pronounced directionalism, both from crystal perfection and from crystal imperfection; and according to the theory advanced earlier in this study,[3] the present authors prefer the interpretation that these formations relate to original crystal growth in which clusters of molecules—"micelles"—have assumed a common orientation to provide the single grain, but have carried into the solid state certain aspects of their original micellar form and a general imperfection pattern of their cumulative mismatching. It is suggested that the grain under observation in Fig. 2 has grown from the lower right as the result of a commonly oriented accumulation of two major lineages and several minor ones. This growth then possibly halted along the observed boundary during solidification, beginning again with renewed nucleation, as indicated in the pattern. On the other hand, descriptions given by Kies, Sullivan, and Irwin are impressive in suggesting that at least a portion of this pattern might be Type II, the indicated nucleations being nucleations of fracture with or without further impression of imperfection structure.

In Fig. 3 a particularly puzzling pattern appears in which a boundary of the type in the previous figure delineates with sharp contrasts the upper and lower portions of the field. It is possible that elaborately complex stress formations during fracture approached simultaneously from above and below, meeting at the indicated boundary to form the observed pattern.

On the other hand, it is submitted as the preferred interpretation here that an unusual growth condition is being observed in which dendritic formations, separately nucleated in the liquid phase and growing with lineage characteristics, have approached from solidification directions above and below in the fractograph, the two masses mutually exhausting the remanent liquid phase as solidification progressed, producing thereby both the boundary and the branching fractographic forms.

* Investigation conducted in the senior author's Laboratory under contract with the ONR.

[1] Kies, Sullivan, and Irwin, J. App. Phys. 21, 716–720 (1950).
[2] C. A. Zapffe and C. O. Worden, Acta Cryst. 2, Part 6, 377–82 (1949).
[3] C. A. Zapffe, Am. Soc. Metals Trans. 42, 387–98 (1950).

Note on "Interference of Growing Spherical Precipitate Particles"

HERSHEL MARKOVITZ
Mellon Institute of Industrial Research, University of Pittsburgh, Pittsburgh, Pennsylvania
July 28, 1950

IN a recent article Wert and Zener[1] derived the differential equation

$$\tau dW/dt=(3/2)(1-W)W^{1/3}$$

to which a solution was desired with the condition that at $t=0$, $W=0$. An analytical integration of this equation is

$$\frac{t}{\tau}=\frac{2}{3}\left[\frac{1}{2}\ln\left(\frac{1+x+x^2}{(1-x)^2}\right)+\sqrt{3}\tan^{-1}\left(\frac{-x\sqrt{3}}{2+x}\right)\right],$$

where $W=x^3$. If various values of W are substituted, corresponding values of t can be obtained. As far as one can tell from their Fig. 3, this solution seems to lead to somewhat better agreement with the experimental results than the numerical integration performed by the authors.

[1] C. Wert and C. Zener, J. App. Phys. 21, 5 (1950).

Journal
of
Applied Physics

Volume 21, Number 12 December, 1950

Electron Microscopy of Wet Biological Tissues by Replica Techniques*

IRWIN W. FISCHBEIN
School of Forestry, University of California, Berkeley, California
(Received April 20, 1950)

The study of wet biological tissues *in situ* using the electron microscope was made possible by modifying existing replica techniques and developing a new wet replication method. The shortcomings of existing replica techniques are discussed and the modifications necessitated are described.

Replicas of wet tissues were produced in extremely thin polystyrene films which could be observed directly in the electron microscope; however, superior results were obtained by preparing silica replicas from the replicated polystyrene films. A modified two-replica silica method proved satisfactory but the three-replica silica method developed was preferable. Electron micrographs of wet wood fiber surfaces used to illustrate these techniques reveal new data concerning submicroscopic cell wall structure.

The problem of interpreting replicas of partially degraded and disorganized materials is illustrated and discussed briefly.

INTRODUCTION

THE successful application of a replica method for the electron microscopy of wet biological structures has not been reported although several replica techniques for dry organic materials have been described.[1–4] While some features of cell wall organization have been revealed by the application of replica methods to dry plant fibers, results in general have been quite unsatisfactory.[1,2] This was the experience in the present study after a survey of existing replica techniques. Failure to obtain satisfactory replicas was attributed in part to the plant tissue being thoroughly dry before embedding took place because plant cell walls are saturated with water in the natural state. Furthermore, it was reasoned that micellar systems in the swollen state would reveal more of their submicroscopic structural arrangement since the microinterstices are distended with water. Therefore, it was necessary to develop an effective wet replication method and at the same time to modify certain aspects of existing replica techniques.

SHORTCOMINGS OF EXISTING REPLICA TECHNIQUES

Gold-shadowed replicas of dry wood fiber surfaces formed by embedding with heat and pressure in thin polystyrene films,[3] or by casting thin films from Collodion[4] and Formvar[5] solutions, failed to show adequate fine structure and contrast. The exact reason for the thin-film replicas failing to show the desired fine structure and contrast is not clear. It may be due in part to improper gold-shadowing technique since varying the thin-film thickness did not affect the contrast conditions.

Because of the excellent contrast generally obtainable with silica replicas, the low pressure silica replica method developed by Barnes, Burton and Scott for the study of organic surfaces was tried.[1] In applying this method, the tissue surface was completely desiccated before replication took place. In addition, excessive destruction of the silica replica occurred during the washing procedure. This probably resulted from the prolonged and vigorous washing necessary for the complete dissolution of the 1-mil polystyrene film specified in the method. In the cases where intact replica film areas were obtained and observed in the electron microscope, the silica replicas failed to show

* From a Ph.D. dissertation submitted in Biophysics at the University of California (January, 1950).

[1] Barnes, Burton, and Scott, J. App. Phys. 16, 730 (1945).

[2] S. F. Kern, J. Polymer Sci. 1, 259 (1946).

[3] M. Swerdlow and G. S. Seeman, Bur. Standards J. Res. 41, 231 (1948), RP1921.

[4] C. W. Hock, Textile Res. J. 18, 366 (1948).

[5] V. J. Schaefer, J. App. Phys. 13, 427 (1942).

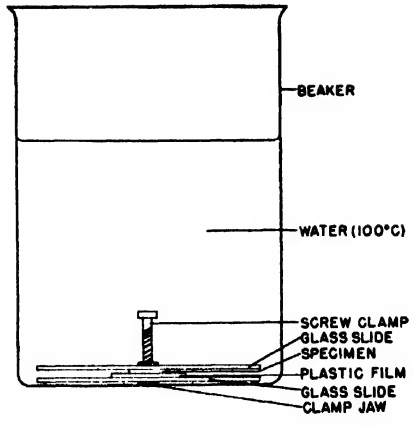


FIG. 1. Sandwiched replica unit with pressure applied submerged in a beaker of hot distilled water for wet replication.

the degree of fine structure to be expected but the contrast conditions were excellent.

The problem of desiccation of the specimen during replication was obviated by the introduction of a new wet replication method. The excessive destruction of the silica replicas during washing was circumvented by the use of extremely thin polystyrene films in the washing procedure and by the introduction of a different solvent and washing apparatus.

Certain shortcomings of the modified two-replica silica method developed in this study were overcome by the use of a third replica. The resulting three-replica silica method was found to be superior from the standpoint of ease of manipulation and of revealing fine structure.

THE WET REPLICATION METHOD

Satisfactory replicas of wood microtome sections, wood fibers, cambial cells and bacterial cells in the wet state were obtained by the following method:

1. The wet object to be replicated is placed on the replicating plastic which is sandwiched between two clean glass microscope slides. A few drops of distilled water are placed on the object before applying the upper glass slide.

2. Pressure is applied directly over the object by means of a medium-sized pinch clamp or Hoffman screw clamp.

3. The sandwiched unit with the pressure applied is completely submerged in a beaker of distilled water at 100°C. (See Fig. 1.) The correct amount of pressure to be applied by the screw clamp is readily attainable in a few trials. When too much pressure is applied, the glass slides crack during heating. The correct pressure is slightly less than that which will cause the glass slides to crack. If a medium-sized pinch clamp is used, this problem does not arise; however, the screw clamp is much easier to manipulate and gives more uniform pressure. The unit remains in the boiling water for approximately 7 minutes when polystyrene or polyethylene are the replicating plastics. The unit is then removed and allowed to cool to room temperature before releasing the pressure clamp.

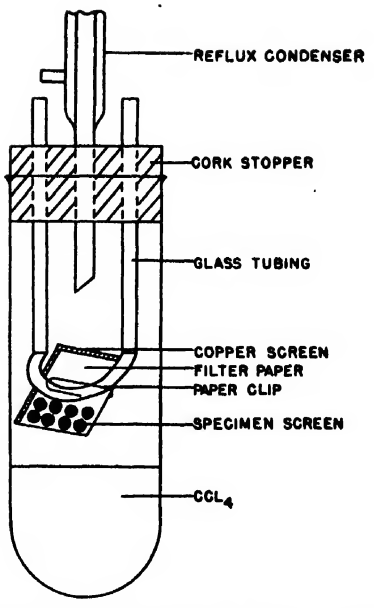


FIG. 2. Apparatus used to wash the polystyrene film from the silica replicas by solvent condensation.

4. The object is allowed to dry and is removed from the plastic replica by mechanical stripping with a tweezer or with a readily adhering pressure sensitive tape.

Although only the use of polystyrene and polyethylene are described in this paper for the initial replica, any plastic with a similar softening point and resistance to water absorption could be used.

A thin-film polystyrene replica made by the wet method in a polystyrene film cast on a glass slide from a 1 percent polystyrene solution in benzene showed more fine structure when gold-shadowed than one made by the dry method. However, the contrast and fine structure observed were much less than that obtainable by the silica methods.

MODIFIED TWO-REPLICA SILICA METHOD

The standard two-replica silica method[6] was modified in the following manner in order to obtain satisfactory silica replicas of wet tissues:

1. An extremely tenuous film is obtained on a clean glass microscope slide by dipping the slide into a 1 percent solution of polystyrene in benzene and allowing the slide to dry while held in a vertical position on a piece of paper toweling. A replica is produced in this thin film by the wet replication method.

2. The thin-film replica is cut into $\frac{1}{8}$-inch squares on the microscope slide with a scalpel or sharp knife and the replica squares are floated from the slide onto a

[6] C. H. Gerould, J. App. Phys. 18, 333 (1947).

water surface by slowly inserting the glass slide at a low angle into a dish of distilled water. The floating replica squares are lifted from the water surface by raising the specimen screens up under them. The specimen screens are then placed on a piece of paper toweling so that the polystyrene film will be pulled down tight and flat on the screen as the water is imbibed by the paper. These thin-film replicas are ready for silica evaporation. (If it is desired, a suitable metal such as gold, chromium, or uranium can be used to shadowcast this replica[7] which can then be observed directly in the electron microscope.)

3. The thin-film replicas mounted on specimen screens are placed in a high vacuum chamber directly above a tightly wound 15° cone-shaped coil filament of 20-mil tungsten wire for silica evaporation. The turns of the tungsten coil must be close together to prevent excess evaporation of the silica through the sides and bottom of the coil. The quantity of silica (2–3 mg) to be evaporated can be conveniently measured into the coil with a micropipette calibrated to deliver the correct volume of Ludox.† The specimen screens are supported in the inverted position above the coil with a piece of transparent cellulose adhesive tape mounted on a glass slide with the adhesive side exposed. Specimen screens readily stick to the tape when pressed lightly against the adhesive surface with a forceps. Transparent adhesive tape used in this manner is a convenient way to carry, catalogue, and store finished specimen screens.

4. After silica evaporation, the double replica mounted on the specimen screen is placed in the washing apparatus shown in Fig. 2. The specimen screens rest on the filter paper square (Whatman No. 1) and are washed gently in a liquid film formed by the condensing vapors after the filter paper is saturated. The washing time and rate are determined empirically and the heating is accomplished by means of a water bath. Only the solvent vapors which condense on the water-cooled glass condensing tube wet the filter paper. The reflux condenser is placed in the system to prevent the danger of a closed system.

After the washing schedule has been established, many finished specimen screens contain a large number of squares covered by the clean intact silica replicas. Carbon tetrachloride was superior to ethyl bromide as the washing agent in this apparatus. This may be due to the fact that carbon tetrachloride does not swell the polystyrene before dissolving it and that the washing can be more readily controlled because carbon tetrachloride is not as volatile as ethyl bromide. After the washing is completed, the filter paper is allowed to dry before the specimen screens are removed to be placed on the adhesive-taped glass slides. These clean silica replicas are ready for observation in the electron microscope.

[7] R. C. Williams and R. C. Backus, J. App. Phys. **20**, 98 (1949).

† Ludox, 30 percent colloidal silica solution in water. Technical Grade. DuPont de Nemours Company, Baltimore, Md.

The completely clean silica replicas prepared in this manner do not display the fragility reported by Heinmets.[8] They can be observed at maximum intensities in the RCA Type EMB and EMU scopes without excessive destruction. However, since the silica films are extremely thin, maximum intensities are not necessary for careful examination. Figure 3 shows a silica replica, made in this manner, of the lumen wall surface of a tracheid of ponderosa pine which has been partially disrupted by drastic chemical treatment. Distinct fibrils which have resulted from this chemical action can be seen grading down to the limit of resolution of the instrument.

Although this method is entirely satisfactory as far as the structures revealed are concerned, the manipulation of thin-film replicas formed by pressure embedding is quite difficult. Many excellent areas are repeatedly lost during the removal of the replicated object from the thin-film polystyrene and also when the thin-film squares are floated onto the water surface. Furthermore, uniform embedding in the thin polystyrene film is very difficult because the film cast from the 1 percent solution is extremely tenuous. The result of this is that the areas most deeply embedded and, therefore, best

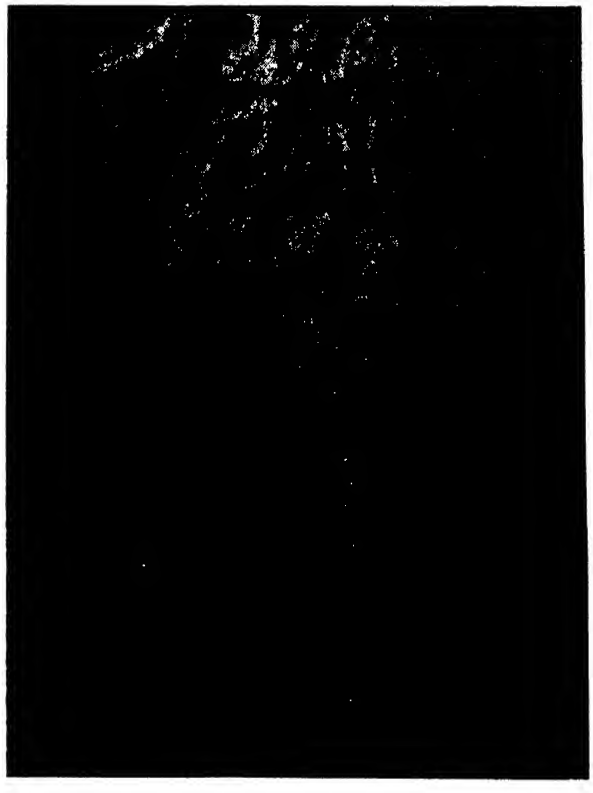


FIG. 3. Surface of the inner layer of the secondary wall of a ponderosa pine fiber after vigorous chemical treatment, showing disorganization of the fibrillar orientation. Silica replica made by the modified two-replica silica method in which the first step is a polystyrene replica of the wet tissue. ×24,000.

[8] F. Heinmets, J. App. Phys. **20**, 384 (1949).

for revealing fine structure are also the most fragile. In an effort to overcome these difficulties, another method was developed in which an extra replication step was introduced. This resulted in simpler, surer manipulation and in a deeper replica of better contrast.

THE THREE-REPLICA SILICA METHOD

The three-replica silica method is a combination of the advantages of a deep replica with adequate embedding and a uniform thin-film replica with ease in washing away the polystyrene.

The steps in the preparation of the final silica replica are as follows:

1. A 4-mils thick film of polyethylene 1-inch square is used for the initial replica, with the wet object replicated in the manner described under the wet replication method.

2. The polyethylene film, after the object has been replicated, is glued to a glass slide with ordinary mucilage. After the mucilage has dried, the object is removed from the polyethylene by teasing it out with tweezers or by stripping it with pressure sensitive tape.

3. The slide with the replicated polyethylene glued to it is immersed in a coplin jar of a 1 percent solution of polystyrene in benzene. The solution is allowed to dry while the slide is held in a vertical position on a piece of paper toweling. This produces a thin, uniform polystyrene film on the replicated polyethylene sheet.

4. The thin-film polystyrene replica is stripped from the polyethylene replica by backing the polystyrene with a thick layer of a 20 percent solution of gelatin in water.[9] After the gelatin has dried, the thin-film polystyrene replica is removed with the gelatin when it is flaked off.

5. The gelatin with the thin polystyrene replica adhering to it is cut into $\frac{1}{8}$-inch squares which are placed gelatin side down into a Syracuse watch glass of hot distilled water. The gelatin dissolves, leaving the small replica squares floating on the water surface with the replica side up. These are then transferred to a large dish of distilled water by means of a piece of 80-mesh screen.

6. The thin-film replicas are lifted from the water surface on the specimen screens which are placed on a piece of paper toweling and allowed to dry. The dry replicas are placed in the vacuum chamber for silica evaporation.

7. The polystyrene is removed by reflux washing in the washing apparatus, and the clean silica films on the specimen screens are ready for electron microscopy.

An example of the results obtainable by this method is shown in Fig. 4 which is a replica of the inner layer of the secondary wall of a redwood tracheid. The low slope of fibrillar orientation of this layer of the wall adjoining the lumen is clearly discernible. This condition of low fibrillar slope in the inner layer of the secondary wall has been observed in all normal coniferous tracheids examined and although this condition had been previously postulated, this is the first time it has been directly demonstrated.

An area from Fig. 4 is shown in Fig. 5 at a higher magnification which makes visible markings on the fibrils. These markings which appear at approximately right angles to the direction of fibrillar orientation may be the result of an alternation of areas with diminishing and increasing crystallinity. The protuberances appearing along the oriented ridges are not necessarily perfect crystalline aggregates. Rather, they can be areas with a greater degree of crystallinity than the intervening areas. If this is the case, it would support the theory put forth by Roseveare and coworkers[10] suggesting that a mechanism exists in crystallizing high polymeric systems for the systematic arrangement of the crystalline aggregates to produce a super-lattice of crystallites. This would explain the fact that in fibrillar cellulose systems planes of weakness are found to exist at right angles to the prevailing direction of fibrillar orientation.[11] The redwood microtome section from which this replica was made was not treated chemically and this is

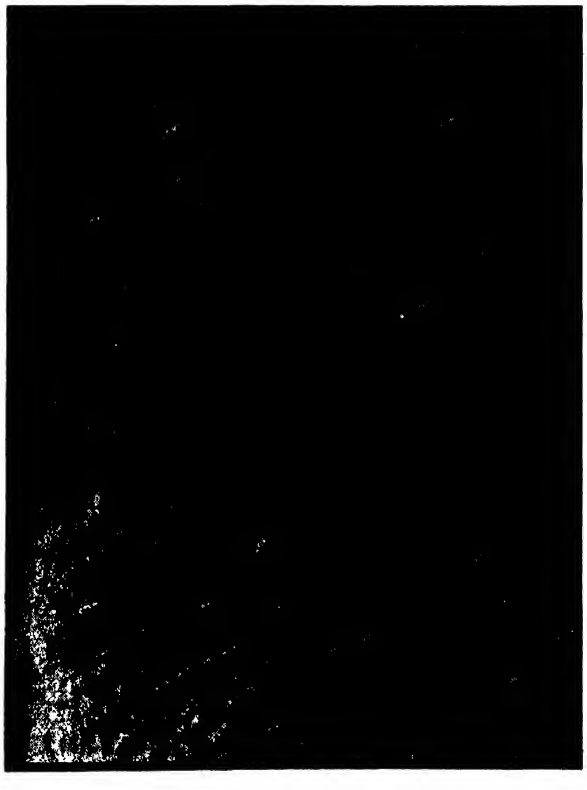


FIG. 4. Surface of the inner layer of the secondary wall of a redwood tracheid without any previous chemical treatment, showing the low slope of fibrillar orientation. The solid line denotes the fiber axis; the dashed line indicates the fibril axis. Silica replica made by the three-replica silica method in which the first step is a polyethylene replica of the wet tissue. ×18,500.

[9] V. J. Schaefer, Phys. Rev. 63, 495 (1942).
[10] Roseveare, Waller, and Wilson, Textile Res. J. 18, 114 (1948).
[11] I. W. Bailey, Bull. Torrey Bot. Club 66, 201 (1939).

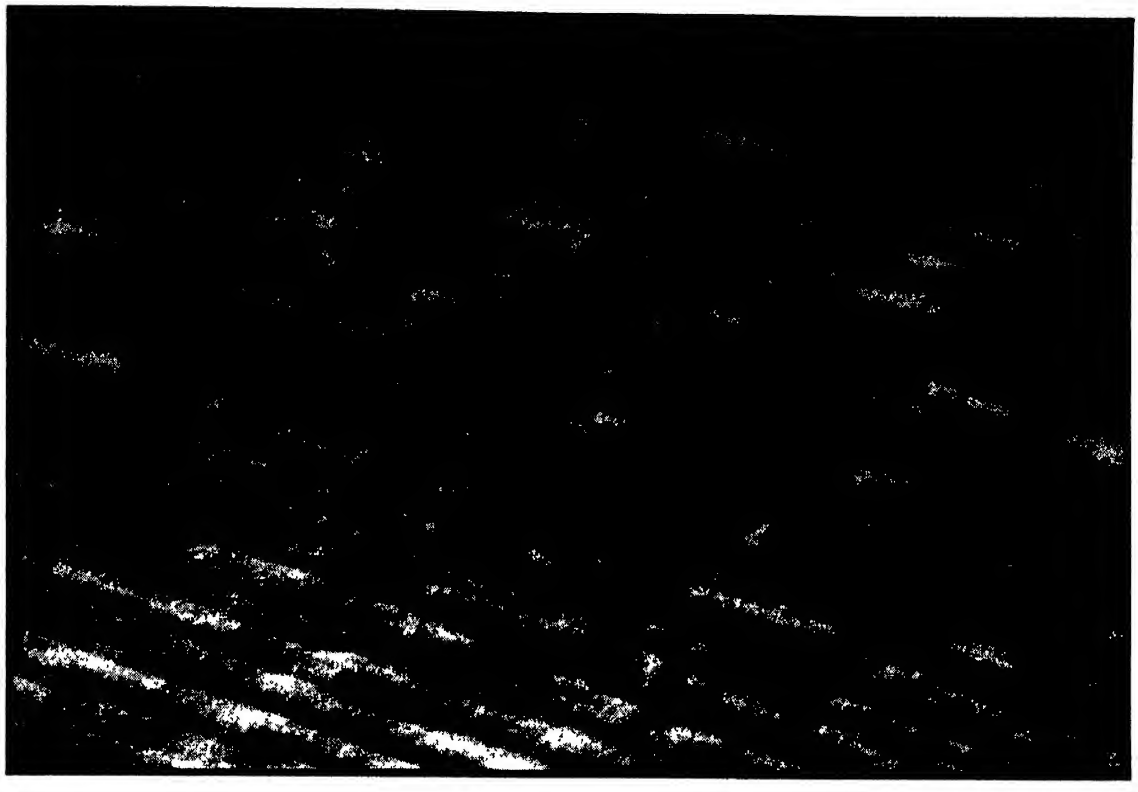


FIG. 5. An area from Fig. 4 highly magnified (×55,000), showing distinct markings on the fibrils.

the first time an unaltered, water-saturated plant fiber wall has been shown by means of the electron microscope.

A replica of a cross section of the walls of ponderosa pine tracheids is shown in Fig. 6. This cross section (20-microns thick) was partially delignified by a holocellulose method before replication took place. The light places in the cross section are areas where lignin has been removed. The striae crossing the wall are due to knife tears produced during sectioning. The middle lamella region consists of the light colored irregular areas, where maximum lignin removal has occurred, separating the walls of adjacent cells.

At the beginning of this study it was assumed that a direct replica method in which the initial replica would be observed in the electron microscope would be preferable to any method in which additional replicas were introduced before the final replica for observation. However, this assumption proved to be incorrect and the three-replica silica method is unquestionably preferable to any other method attempted in this study. Although the three-replica silica method appears to be quite complicated, it is actually possible for an experienced technician to produce a large number of excellent silica replicas in a relatively short time. For example, in this study as many as 50 specimen screens with a large number of observable areas of intact silica replicas could be prepared in a 2-day period.

FIG. 6. Cross section of the walls of partially delignified ponderosa pine fibers. Portions of the walls and lumina of three adjoining cells is shown. Silica replica produced by the three-replica silica method. ×4500.

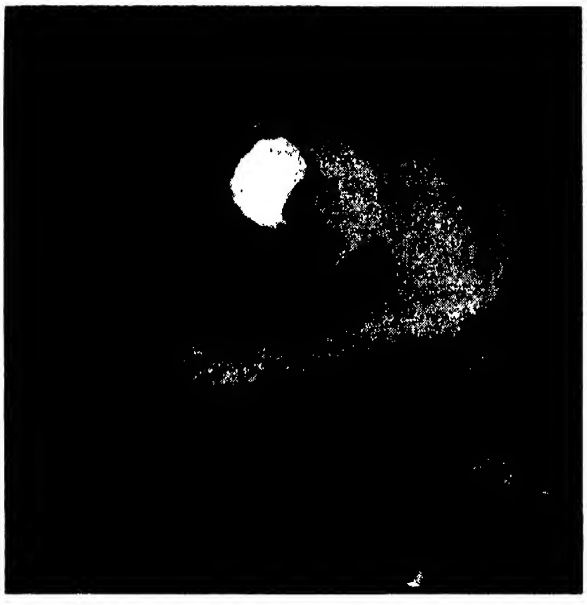


FIG. 7. Bordered pit of ponderosa pine fiber which has undergone drastic chemical treatment showing fibrils reoriented radially around the pit aperture by the replicating procedure. Silica replica produced by the three-replica silica method. ×4000.

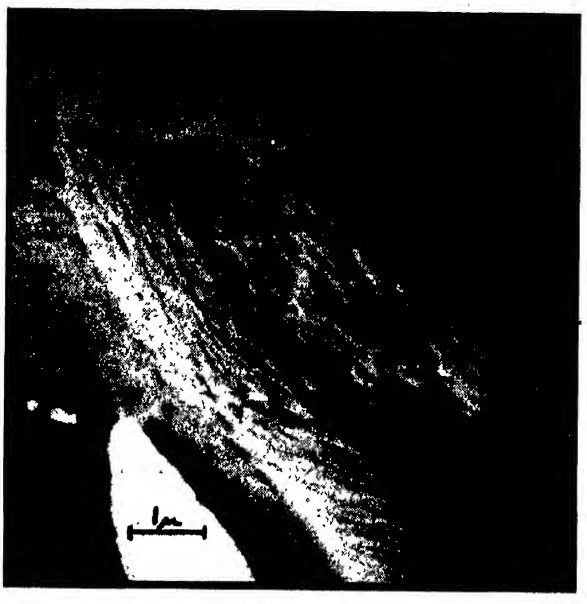


FIG. 8. Bordered pit of ponderosa pine fiber which has undergone very mild chemical treatment showing the true circular fibrillar orientation of the cellulose around the pit aperture. Silica replica produced by the three-replica silica method. ×10,000.

Using the three-replica silica method, it has been possible to use the electron microscope to study the submicroscopic structural relationships of the layers and markings of the secondary cell wall of coniferous tracheids. A later paper will discuss the results of this study.

INTERPRETATION OF REPLICAS

The interpretation of replicas of partially degraded and disorganized materials is a serious problem. In Fig. 7 is shown a replica of a bordered pit in the outer layer of the secondary wall of a fiber produced by a drastic maceration method. Fibrils with an apparent radial orientation around the pit aperture can be seen. These radially oriented fibrils were pulled into this position during the replication process. This interpretation is correct since the bordered pit in this same layer from a replica of a fiber produced by a mild maceration method shown in Fig. 8 displays circular rather than radial orientation around the pit aperture. This is in agreement with the accepted interpretation of bordered pit structure in this layer.

Loosely aggregated substances can be disorganized and reoriented by replicating pressures and, therefore, replicas of such materials should be interpreted with caution.

The author is indebted to Dr. Robert A. Cockrell of the School of Forestry and to Dr. William N. Takahashi of the College of Agriculture, University of California, for their helpful suggestions during the course of the study.

Magneto-Ionic Triple Splitting of Ionospheric Waves

O. E. H. RYDBECK
Research Laboratory of Electronics, Chalmers University of Technology, Gothenburg, Sweden
(Received November 28, 1949)

In recent years reports have appeared from various ionospheric observatories concerning the appearance of a third magneto-ionic component or a so-called z trace. The phenomenon has also been called magneto-ionic triple splitting. After having recorded a number of such triple splits at the new Kiruna observatory, we became very interested in the problem as a whole. The entire matter was examined theoretically and as a result it has been possible to show that strong coupling often exists between the ordinary and z components as a critical level. It has further been possible to show that the z wave becomes strongly excited when the collisional frequency ν approaches the critical collisional frequency ν_c of Appleton and Builder. When ν is larger than ν_c the o echo rapidly disappears and only the z and x components remain (if ν_c is sufficiently small, *i.e.* at very high latitudes). A useful expression for the transmission coefficient of the $o-z$ wave has been derived.

INTRODUCTION

TRIPLE splitting of ionospheric waves or rays was reported in 1933 by Eckersley.[1] A few years later, in 1935 and 1936, Toshniwal[2] and Harang[3] reported similar phenomena. This interesting matter did not attract much attention until after the war the chain of recording ionospheric observatories had been greatly increased. According to Meek,[4] triple splitting was observed in Canada in 1943 when the Churchill ionosphere station commenced operation. Since then the phenomenon was observed frequently at other Canadian stations as well. In Hobart, Tasmania, in 1946, Newstead observed F_2 triple splitting.[5] Seaton, reporting from College, Alaska, in 1947, presents a record of F_2 triple splitting[6] and discusses the possible mechanism of excitation of the third or z component of the downcoming wave. Further, Gipps, Gipps, and Venton at Brisbane[7] as late as August, 1948 describe a ray trace which appears to have the characteristics of the z component. It is significant that all the reports mentioned come from places of high or fairly high geomagnetic latitude.

At the Kiruna Ionospheric and Wave Propagation Observatory, which started regular recording on October 1, 1948,[8] a large number of triple splits have been recorded both in the E and the F regions. As far as is known to us, E triple splitting has only been reported earlier by Meek.[4] The fairly frequent occurrence of ionospheric triple splitting at Kiruna has made us particularly interested in the phenomenon. The purpose of the present note is to make a preliminary investigation of the excitation and properties of this interesting z component of the wave.

THEORETICAL CONSIDERATIONS

In his treatise, "On the propagation of radio waves,"[9] the author has described the coupled wave equations governing the propagation of the ordinary and extraordinary waves in a plane inhomogeneous ionosphere. With the symbols used in that work[9] and the assumption that the electron density varies in the phase propagation or z direction only these equations become

$$\frac{d^2\Pi_1}{dz^2}+(k_0^2\epsilon_1+\psi^2)\Pi_1=-\pi_2\frac{d\psi}{dz}-2\frac{d\Pi_2}{dz}\cdot\psi, \tag{1}$$

$$\frac{d^2\Pi_2}{dz^2}+(k_0^2\epsilon_2+\psi^2)\Pi_2=-\pi_1\frac{d\psi}{dz}-2\frac{d\Pi_1}{dz}\cdot\psi. \tag{2}$$

The values of ϵ_1 and ϵ_2 are given by the Appleton-Hartree formula,[10] $k_0=2\pi/\lambda_0$ (where λ_0 is the vacuum wave-length), and ψ is a coupling coefficient. One further has (if the time factor is $e^{-j\omega t}$)

$$\psi=-\frac{j}{2}\cdot\frac{d(v^2)}{dz}\cdot\frac{\delta_c}{(v^2-ja_1)(v^2+ja_2)}, \tag{3}$$

with

$$\left.\begin{aligned}a_1&=\delta_c+\delta,\\ \text{and}\quad a_2&=\delta_c-\delta.\end{aligned}\right\} \tag{4}$$

In these formulas

$$\delta=\nu/\omega, \tag{5}$$

where ν is the frequency of collision and $\omega=2\pi\cdot f$, the wave angular frequency, is a loss angle measure of the medium. The critical loss angle is

$$\delta_c=\nu_c/\omega, \tag{6}$$

where ν_c is the critical collisional frequency of Appleton-

[1] T. L. Eckersley, "Discussion of the ionosphere," Proc. Roy. Soc. A141, 710 (1933).
[2] G. R. Toshniwal, "Threefold magneto-ionic splitting of the radio echoes reflected from the ionosphere," Nature 135, 471 (1935).
[3] L. Harang, "Vertical movements of air in the upper atmosphere," Terr. Mag. and Atmos. Elec. 41, 160 (1936).
[4] J. H. Meek, "Triple splitting of ionospheric rays," Nature 161, 597 (1948).
[5] G. Newstead, "Triple magneto ionic splitting of rays reflected from the F_2 region," Nature 161, 312 (1948).
[6] S. L. Seaton, "Magnetoionic multiple refraction at high latitudes," Proc. I.R.E. 36, 450 (1948).
[7] Gipps, Gipps, and Venton, "Note on night-time phenomena in the F_2 region at Brisbane," J. Council Sci. Ind. Research 21, 215 (1948).
[8] O. E. H. Rydbeck, "The Ionospheric and Radio Wave Propagation Observatory at Kiruna, 67°50′ N, 20°14.5′ E," Tellus 1, No. 4 (1949).
[9] O. E. H. Rydbeck, 1944. "On the propagation of radio waves," Trans. Chalmers University of Technology, No. 34.
[10] E. V. Appleton, "Wireless studies of the ionosphere," Inst. Elec. Eng. 71 (1932).

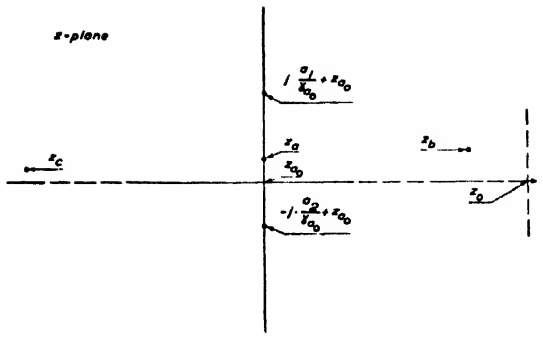


FIG. 1. Showing the location of the zero of n^2 and the poles of ψ in the z plane.

Builder,[11] *viz.*,

$$\nu_c = \tfrac{1}{2}\cdot(\sin^2\theta/\cos\theta)\cdot\omega_H, \tag{7}$$

and θ is the angle between the z axis and the direction of the terrestrial magnetic field. Further $\omega_H = 2\pi\cdot f_H$ is the angular frequency of gyration of the free electron in this field, and

$$\nu^2 = (\omega_c^2/\omega^2) - 1, \tag{8}$$

where ω_c is the ordinary wave critical angular frequency of the electron density level in question. For a parabolic layer formed between $z = z_0 - \Delta h$ and $z = z_0 + \Delta h$, for example, one would have

$$\omega_c^2 = \omega_{cm}^2\{1 - [(z - z_0)/\Delta h]^2\}, \tag{9}$$

where ω_{cm} is the ordinary wave critical angular frequency of the layer.

In the wave equations (1) and (2) we assume that index (1) denotes the ordinary or o component and index (2) the "unusual" z component.

For the unperturbed o wave functions we use the following approximations*

$$\overset{(2)}{\Pi_1}{}^{0(1)} = (\pi W_1/2n_1)^{\frac{1}{2}}\exp(\pm j\pi/6)\cdot\overset{(1)}{H}_{1/3}{}^{(2)}(W_1), \tag{10}$$

where $n_1 = (\epsilon_1)^{\frac{1}{2}}$ is the index of refraction of the o mode and W_1 the phase integral

$$W_1 = \int_{z_a}^{z} k_0 n_1 dz = W_1(z_a, z). \tag{11}$$

It is important to note that z_a is the location of the zero of ϵ_1, i.e., the branch point of n_1. When we have losses, this point (or points) lie in the complex z plane.

Assuming that to the left of the layer $n_1 = e^{-j\pi}$, $\Pi_1{}^{0(1)}$ apparently represents a wave incident upon the layer. Neglecting the possible formation of a z wave we see that the coefficient of reflection, at the lower boundary $z_0 - \Delta h$, becomes

$$R \cong \exp[j\cdot 2\{W_1(z_a, z_0 - \Delta h) - \pi/4\}] \tag{12}$$

if the layer is not very thin and $\omega \neq \omega_{cm}$.

For the unperturbed z wave functions we use similar wave functions denoted $\overset{(2)}{\Pi_2}{}^{0(1)}$ with

$$W_2 = \int_{z_b}^{z} k_0 n_2 dz = W_2(z_b, z), \tag{13}$$

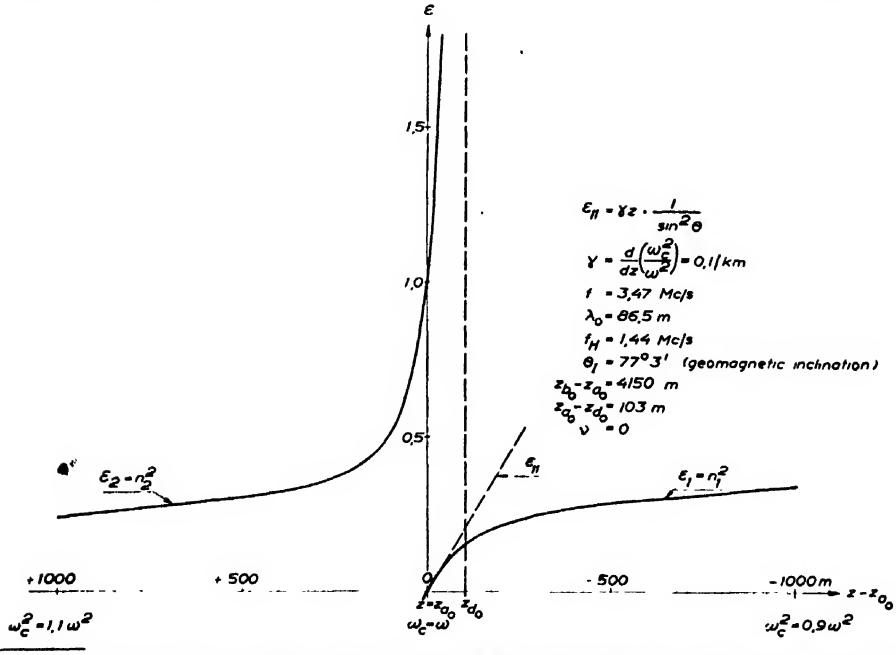


FIG. 2. Index of refraction in the neighborhood of $\omega_c = \omega$ when $\nu = 0$. *Note.* $|\epsilon_2| = \infty$ for $z = z_d$, at which point (branch point of Rai)

$$\frac{\omega_c^2}{\omega^2} = \frac{1 + j\delta}{1 + \dfrac{\gamma^2\sin^2\theta}{(1+j\delta)^2 - \gamma^2}}.$$

[11] E. V. Appleton and G. Builder, "The ionosphere as a doubly refracting medium," Proc. Phys. Soc. 45 (1933).

* Generally sufficiently accurate except in the neighborhood of a second zero of ϵ_1. When two zeros come very close (not of immediate interest in this connection) parabolic wave functions have to be used. See O. E. H. Rydbeck, "On the propagation of waves in an inhomogeneous medium," Report No. 7 from the Research Laboratory of Electronics. Trans. Chalmers University of Technology **74**, 23 (1948).

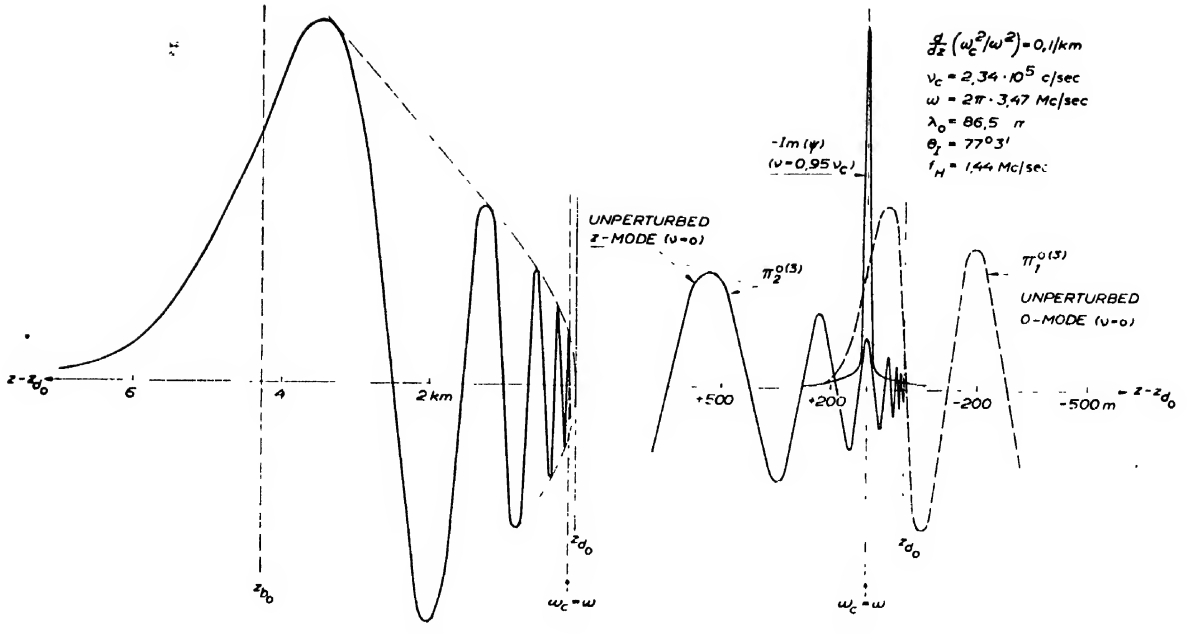


FIG. 3. The coupling coefficient and the unperturbed wave functions in the interaction region.

where z_b is the position of the appropriate ϵ_2 zero. A fundamental question apparently is to study the behavior of n_1 and n_2 in the z plane. To this question we will return briefly after first having studied the nature of the coupling coefficient.

Denoting by $z=z_{a0}$ the interaction level where $\omega_c^2=\omega^2$ (also branch point of n_1 when $\nu=0$), we use the following approximation in the neighborhood of this level, *viz.*,

$$\nu^2=\left\{\frac{d}{dz}\left(\frac{\omega_c^2}{\omega^2}\right)\right\}_{z_{a0}}(z-z_{a0})=\gamma_{a0}(z-z_{a0}). \tag{14}$$

Under these circumstances we have

$$\psi=-\frac{j}{2}\cdot\frac{\delta_c/\gamma_{a0}}{\{(z-z_{a0})-ja_1/\gamma_{a0}\}\{(z-z_{a0})+ja_2/\gamma_{a0}\}}. \tag{15}$$

It is extremely interesting, from an interaction point of view to find, that

$$\int_{-\infty}^{+\infty}\psi dz=\begin{cases}0, & \nu>\nu_c\\ -j\pi/2, & \nu<\nu_c\end{cases}. \tag{16}$$

Returning now to the question of the refractive index, we find it convenient to use the following form of the Appleton-Hartree formula, *viz.*,

$$n_{\frac{1}{2}}^2=1-\frac{(\nu^2+1)(-\nu^2+j\delta)}{(1+j\delta)(-\nu^2+j\delta)-\cos\theta\cdot y\{\delta_c\mp[\delta_c^2+(-\nu^2+j\delta)^2]^{\frac{1}{2}}}. \tag{17}$$

One easily verifies that n_1^2 is zero for $\nu^2=j\delta$, and n_2^2 zero for $\nu^2=\pm y+j\delta$, where $y=\omega_H/\omega$. Thus, even considering the losses the zeros of n_1^2 and n_2^2 are independent of θ.

Denoting by z_c the position of the n_2^2 zero corresponding to $\nu^2=-y+j\delta$, we find for the parabolic layer

$$z_a\cong z_0-\Delta h\left[1-\left(\frac{\omega}{\omega_{cm}}\right)^2\right]^{\frac{1}{2}}+j\frac{\delta\cdot\Delta h}{2}\cdot\frac{\omega/\omega_{cm}}{[(\omega_{cm}/\omega)^2-1]^{\frac{1}{2}}}(\omega^2<\omega_{cm}^2), \tag{18a}$$

$$z_b\cong z_0-\Delta h\left[1-\left(\frac{\omega}{\omega_{cms}}\right)^2\right]^{\frac{1}{2}}+j\cdot\frac{\delta\Delta h}{2}\cdot\frac{\omega^2/\omega_{cm}^2}{[1-(\omega/\omega_{cms})^2]^{\frac{1}{2}}}\quad[\omega^2<\omega_{cms}^2=\omega_{cm}^2/(1+y)], \tag{18b}$$

and

$$z_c\cong z_0-\Delta h[1-(\omega/\omega_{cmz})^2]^{\frac{1}{2}}+j\frac{\delta\Delta h}{2}\cdot\frac{\omega^2/\omega_{cm}^2}{[1-(\omega/\omega_{cmz})^2]^{\frac{1}{2}}}\quad[\omega^2<\omega_{cmz}^2=\omega_{cm}^2/(1-y)]. \tag{18c}$$

Thus,

$$\mathrm{Im}(z_b)>\mathrm{Im}(z_a)>\mathrm{Im}(z_c).$$

It is especially interesting to find that

$$\mathrm{Im}(z_a)=(a_1-a_2)/2\gamma_{a0}, \tag{19}$$

$$\frac{1}{\gamma_{a0}}=\frac{\Delta h}{2}\cdot\frac{\omega/\omega_{cm}}{[\omega_{cm}^2/\omega^2-1]^{\frac{1}{2}}}.$$

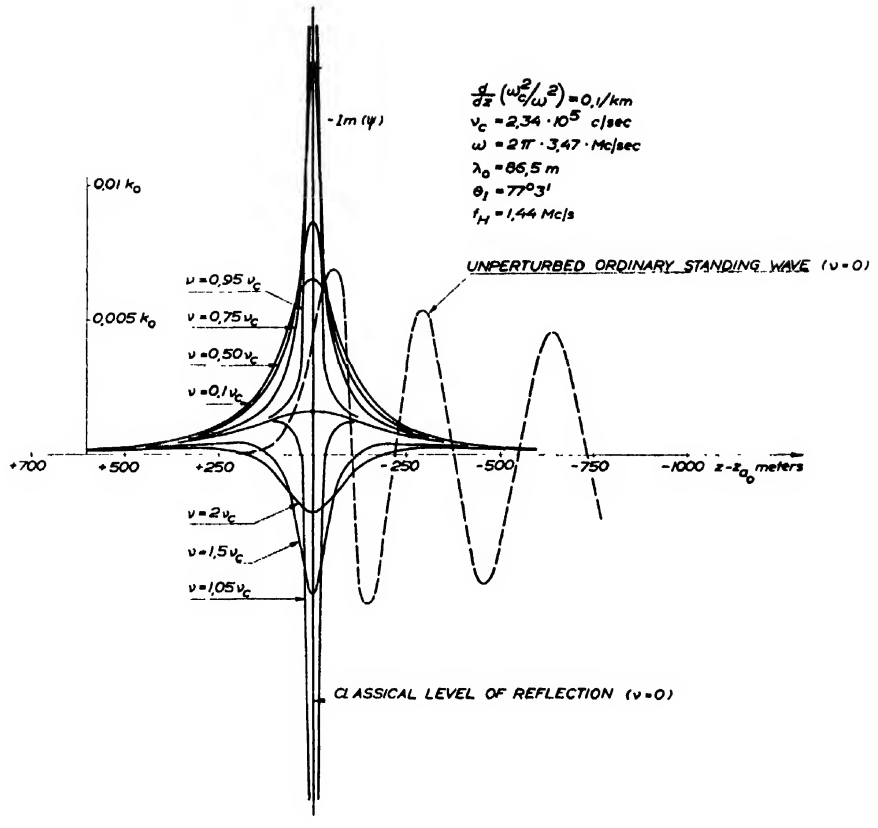


FIG. 4. Demonstrating the nature of $-\mathrm{Im}(\psi)$.

As $\mathrm{Re}(z_a) = z_{a0}$ this means that the zero of n_1^2 is located midway between the poles $ja_1/\gamma_{a0} + z_{a0}$ and $-ja_2/\gamma_{a0} + z_{a0}$ of ψ. The situation is depicted in Fig. 1.

It is further interesting to note that

$$(n_2)_{z=z_a} = 1 \tag{20}$$

and that at the poles $n_1 = n_2$.

We have in Fig. 2 plotted the variation of ϵ_1 and ϵ_2 for $\nu = 0$, a normal electron density gradient and geomagnetic conditions of the Kiruna observatory. The unperturbed standing wave functions,

$$\Pi_1^{0(3)} = \Pi_1^{0(1)} + \Pi_1^{0(2)} \quad \text{and} \quad \Pi_2^{0(3)} = \Pi_2^{0(1)} + \Pi_2^{0(2)},$$

corresponding to these ϵ-variations are shown in Fig. 3 for $\nu = 0$ together with $-\mathrm{Im}(\psi)$ for $\nu = 0.95\nu_c$. This figure demonstrates that the coupling actually takes place within a very narrow region of a few hundred meters effective extension.

It is of particular interest in this connexion to demonstrate the peculiar nature of ψ for various ν-values. The variation of the important component $-\mathrm{Im}(\psi)$ is therefore depicted in Fig. 4 from which it is easy to understand that the "interaction integral" (16) is zero when $\nu > \nu_c$.

Let us next study the first-order approximations to solutions of the excited z waves, running up and down between z_d and z_b. Introducing

$$\xi_1^{(1)}(z) = -\Pi_1^{0(1)} \cdot \frac{d\psi}{dz} - 2 \cdot \frac{d\Pi_1^{0(1)}}{dz} \cdot \psi, \text{ etc.,} \tag{21}$$

we find the first-order z wave approximations

$$\Pi_2^{(1)} = -\Pi_2^{0(1)} \cdot \frac{1}{w_2^0} \cdot \int_{b_1}^{z} \Pi_2^{0(2)} (\xi_1^{(1)} + \xi_1^{(2)}) dz \tag{22}$$

and

$$\Pi_2^{(2)} = \Pi_2^{0(2)} \cdot \frac{1}{w_2^0} \cdot \int_{b_2}^{z} \Pi_2^{0(1)} (\xi_1^{(1)} + \xi_1^{(2)}) dz. \tag{23}$$

Here b_1 and b_2 are lower and upper boundaries of the effective interaction region. Further, w_2^0 is the Wronskian of the pair $\Pi_2^{0(1)}$, $\Pi_2^{0(2)}$. With our form of functions $w_2^0 = j \cdot 2k_0$.

Of practical interest for the moment only is to study $\Pi_2^{(1)}$ when $z > b_2$ and $\Pi_2^{(2)}$ when $z < b_1$, i.e., we are not going to study in detail how the Π_2-functions build up within the interaction zone.

At first sight it appears as if one of the two terms of (21) would complicate the treatments of the interaction integrals (22) and (23). On the contrary one finds, after partial integration, that the presence of both these terms makes it possible to analyze the interaction integrals fairly simply.

We thus find for $z>b_2$:†

$$\Pi_2^{(1)} = -\Pi_2^{0(1)} \cdot \frac{1}{w_2^0} \cdot \int_{b_1}^{b_2} \left\{ \Pi_1^{0(2)} \cdot \Pi_2^{0(2)} \left(\frac{\Pi_1^{0(2)'}}{\Pi_1^{0(2)}} - \frac{\Pi_2^{0(2)'}}{\Pi_2^{0(2)}} \right) \psi + \Pi_1^{0(1)} \cdot \Pi_2^{0(2)} \left(\frac{\Pi_1^{0(1)'}}{\Pi_1^{0(1)}} - \frac{\Pi_2^{0(2)'}}{\Pi_2^{0(2)}} \right) \psi \right\} dz, \quad (24)$$

and similarly for $z<b_1$

$$\Pi_2^{(2)} = \Pi_2^{0(2)} \cdot \frac{1}{w_2^0} \times \int_{b_2}^{b_1} \left\{ \Pi_1^{0(2)} \cdot \Pi_2^{0(1)} \left(\frac{\Pi_1^{0(2)'}}{\Pi_1^{0(2)}} - \frac{\Pi_2^{0(1)'}}{\Pi_2^{0(1)}} \right) \psi + \Pi_1^{0(1)} \cdot \Pi_2^{0(1)} \left(\frac{\Pi_1^{0(1)'}}{\Pi_1^{0(1)}} - \frac{\Pi_2^{0(1)'}}{\Pi_2^{0(1)}} \right) \psi \right\} dz. \quad (25)$$

We are of course principally interested in Eq. (24) which yields the amplitude of the wave propagating toward z_b. As far as the interaction integral is concerned replacing b_1 by $-\infty$ and b_2 by $+\infty$ does not appreciably affect its value if b_1 and b_2 are well outside the main interaction zone. We thus write

$$\underset{z>b_1}{\Pi_2^{(1)}} = -\Pi_2^{0(1)} \cdot \frac{1}{w_2^0} \cdot J_2^{(1)}(b_1, b_2) \cong -\Pi_2^{0(1)} \cdot \frac{1}{w_2^0} \cdot J_2^{(1)}(-\infty, +\infty). \quad (24a)$$

If the path of integration is displaced to encircle the pole z_2 of ψ, it is possible to show that the dominant contribution to $J_2^{(1)}$ actually comes from the pole. This pole also is a branch point of n_1 and n_2 (though not a zero point), and in the first order approximation we may still use the approximate wave functions (10).

We have already mentioned that at the ψ-poles $n_1=n_2$. The distance between the poles and z_a, δ_c/γ_{a0}, is quite large. For the locality in question (Kiruna) $\nu_c \cong 2.1 \cdot 10^5$ sec.$^{-1}$ and with the gradient chosen this distance is greater than about 100 meters. The amplitude of W_1 at the poles therefore will be at least of the order 3, the corresponding phase angle with the chosen orientation somewhere between $-\pi/2$ and $-3\pi/4$ at $z_{a0}+j\cdot a_1/\gamma_{a0}$ and between $+\pi/2$ and $+3\pi/4$ at $z_{a0}-ja_2/\gamma_{a0}$. For $W_2[z_{a0}, \pm j\cdot(a_2'/\gamma_{a0})+z_{a0}]$ the amplitude is about three times as large with phase angles approximately $\mp\pi/2$. We can therefore use the asymp-

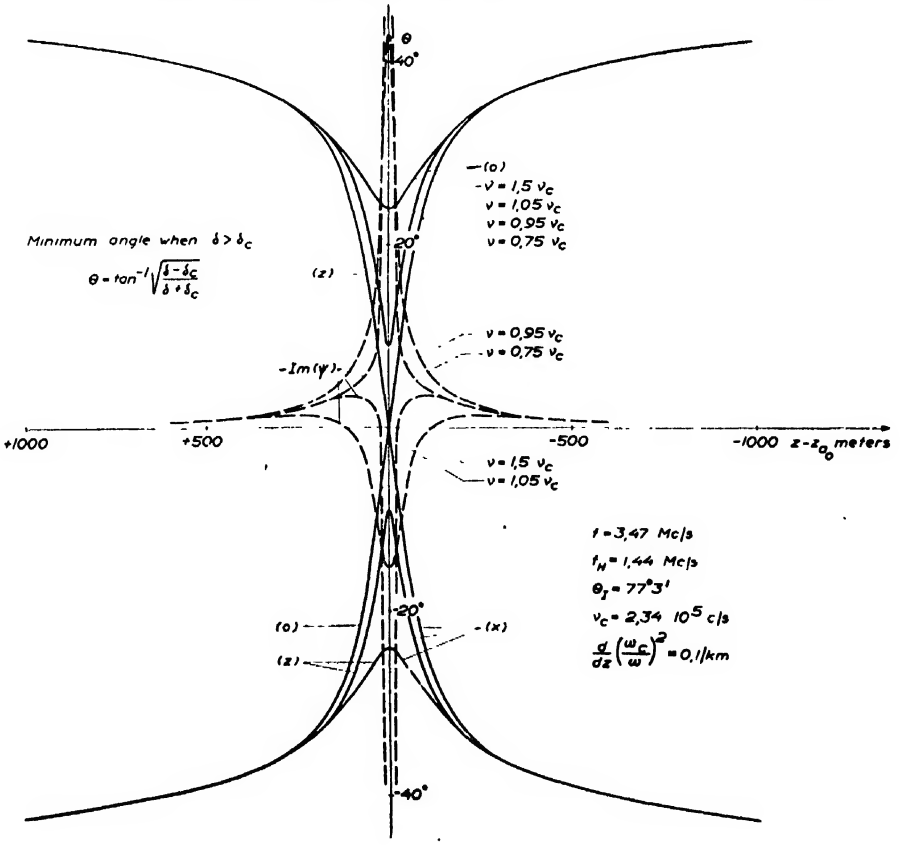


FIG. 5. Depicting the variation in polarization in the coupling region.

† *Note.* $\Pi_1^{0(2)'} = d\Pi_1^{0(2)}/dz$, etc.

totic representation of the Hankel functions at the poles and as there $n_1=n_2$ we infer that the first term of $J_2^{(1)}$ practically disappears (the parenthesis tends to zero). Moreover, $|\Pi_1^{0(2)}|\ll|\Pi_1^{0(1)}|$ at the lower pole. It can therefore safely be said that the first term of $J_2^{(1)}$ does not contribute to its value. We have thus shown that only the "incident" wave function produces the z wave proceeding in the same direction. This is only virtual as the $\Pi_1^{0(2)}$ function must be present to make the integral converge over the path. Also from a physical point of view $\Pi_1^{0(1)}$ cannot be present without $\Pi_1^{0(2)}$.

For the $\Pi_2^{(2)}$-component the result is similar. This time, however, we have to evaluate the residue at $z_{a0}+j\cdot a_1/\gamma_{a0}$ as the wave function products are bounded only in the upper half-plane. Virtually only the "reflected" wave function $\Pi_1^{0(2)}$ contributes to the residue and therefore to the $\Pi_2^{(2)}$-component.

As at the pole

$$\frac{\Pi_1^{0(1)\prime}}{\Pi_1^{0(1)}}-\frac{\Pi_2^{0(2)\prime}}{\Pi_2^{0(2)}}\cong 2\cdot jk_0\cdot n_{\text{pole}},$$

we find

$$\Pi_{2\,z>b_2}^{(1)}(z)\cong\Pi_2^{0(1)}\cdot\exp\Big[-j\Big\{W_1\Big(z_{a0},-j\cdot\frac{a_2}{\gamma_{a0}}+z_{a0}\Big)-W_2\Big(z_{a0},-j\frac{a_2}{\gamma_{a0}}+z_{a0}\Big)\Big\}\Big]\cdot\pi/2\exp(-j\pi/2).$$

Actually we should have started with unit amplitude value of the wave incident at $z=z_0-\Delta h$. The $\Pi_1^{0(1)}(z)$ wave function of the interaction integral therefore should have been replaced by $\Pi_1^{0(1)}(z)/\Pi_1^{0(1)}(z_0-\Delta h)$. Denoting the pole positions by z_1 and z_2, respectively, and introducing this scaled-down wave function we find, if $z>b_2$ but $z\neq z_b$ that

$$\Pi_2^{(1)}(z)\cong\frac{\pi}{2}\cdot\frac{1}{(n_2)^{\frac{1}{2}}}\cdot\exp[j\{W_1(z_{a0},z_0-\Delta h)+W_2(z,z_{a0})\}]\cdot\exp[j\{W_2(z_{a0},z_2)-W_1(z_{a0},z_2)-\pi/2\}]. \quad (26)$$

As the first exponential is the straight phase integral up to z (though with exchanged refractive index after the z wave "takes over") we are led to introduce the first order transmission coefficient T^+ of direction into the layer, *viz.*,

$$T^+\cong(\pi/2)\cdot\exp[j\{W_2(z_{a0},z_2)-W_1(z_{a0},z_2)-\pi/2\}]. \quad \nu<\nu_c. \quad (27)$$

As

$$z_2-z_{a0}=-j\frac{\delta_c-\delta}{\gamma_{a0}}=-j\frac{\nu_c-\nu}{\omega}\cdot\frac{\Delta h}{2}\cdot\frac{\omega/\omega_{cm}}{[(\omega_{cm}^2/\omega^2)-1]^{\frac{1}{2}}},$$

we see that as ν approaches ν_c, $|T^+|$ approaches $\pi/2$.

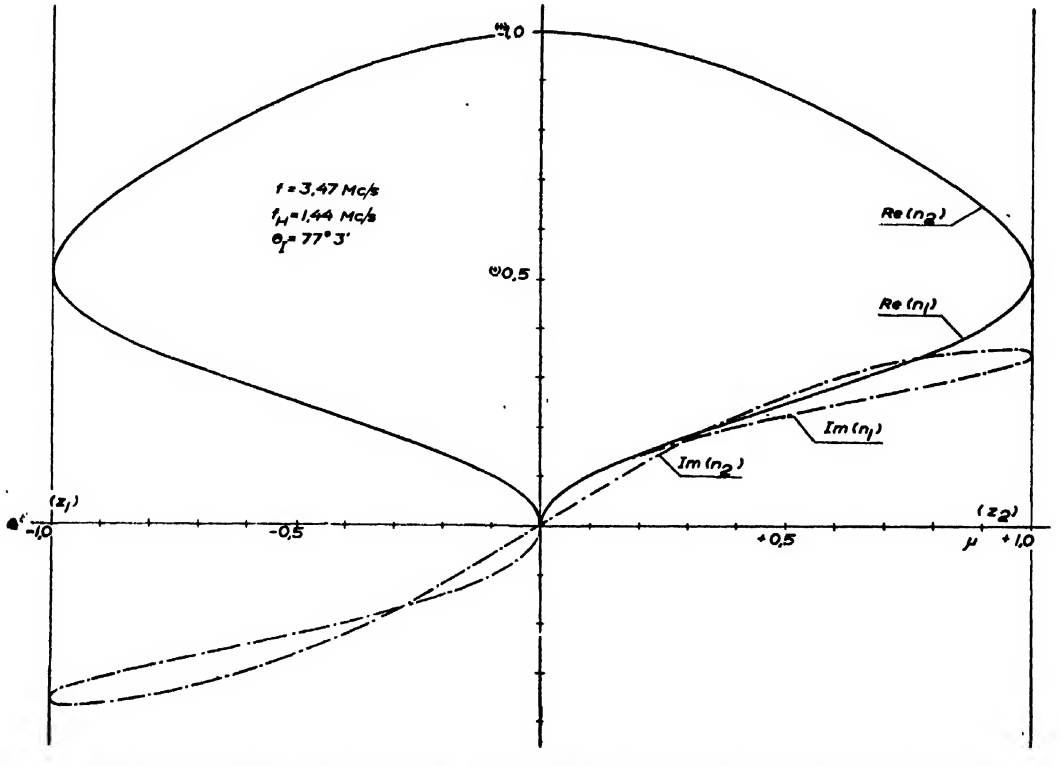


FIG. 6. Showing the variation of the refractive indices for the O^- and Z^- modes along the imaginary axis, joining the ψ-poles Z_1 and Z_2. $[\mu=-(\nu^2-j\delta)/j\delta_c.]$

This indicates very strong coupling, as indeed one would expect. When, namely, ν approaches ν_c the mathematical reflection point, z_a, comes farther and farther away from the interaction point z_{a0}. The interaction pole, z_2, at the same time comes closer and closer to the interaction point and more and more of the incident wave energy is transformed into the z wave. Finally there appears to be complete transformation. However, when the coupling becomes so strong, our first-order approximations cannot be accurate enough. However, if approximations of higher order are also considered, one finds that as ν approaches ν_c the amplitude factor $\pi/2$ of T^+ effectively is reduced to 1.[12] For practical calculations we therefore put

$$T^+\cong\exp[j\{W_2(z_{a0}, z_2)-W_1(z_{a0}, z_2)-\pi/2\}]. \quad \nu\leq\nu_c. \quad (27a)$$

It is obvious that as ν becomes larger than ν_c, z_2 moves into the upper half-plane, and thus in the first-order approximation

$$T^+=0, \quad \nu>\nu_c. \quad (27b)$$

How can this result be explained physically? Let us look at Fig. 5 where we have depicted the polarization of the o and x (z) components as functions of distance in the neighborhood of $\omega_c^2=\omega^2$. The polarization curves only cross each other when $\nu\leqslant\nu_c$ (compare the earlier results of Mary Taylor[13]) and at the level $\omega_c^2=\omega^2$ where the o and z components have the same polarization, in this case plane. At this level the main coupling must occur.

We have further plotted $-\mathrm{Im}(\psi)$ as a function of distance in Fig. 5 in order to show how rapidly the polarization varies through the coupling range when $\nu\leqslant\nu_c$. When ν approaches ν_c the coupling becomes very strong in a narrow region where the difference in polarization is quite small.

As n_1 and n_2 represent the same quantity in different Riemann-surfaces, the cuts joining these surfaces run from z_1 and z_2, say to respectively $v^2=\pm j\infty$. If $\pi_2^{0(1)}$ is "carried" through the cut from the n_2 to the n_1 plane, one immediately infers that along the axis of reals ($z<z_{a0}$) it represents an o wave in the latter plane, traveling in the same direction as $\pi_1^{0(1)}$. Obviously the transformation coefficient, in this geometrical interpretation, becomes

$$T_{0z}\cong\exp[j\{W_2(z_{a0}, z_2)-W_1(z_{a0}, z_2)\}]. \quad \nu\leq\nu_c. \quad (27c)$$

When $\nu>\nu_c$ the axis of reals runs through the cut and the o wave automatically transforms into a z wave, *i.e.* we have longitudinal transmission. Therefore

$$T_{0z}\cong 1, \quad \nu\geq\nu_c \quad (28)$$

which, of course, in no way contradicts the result (27b) of the coupling theory.

[12] O. E. H. Rydbeck, "Magneto ionic triple splitting," Trans. Chalmers University of Technology (in press).

[13] Mary Taylor, "The Appleton-Hartree formula and dispersion curves for the propagation of electromagnetic waves through an ionized medium in the presence of an external magnetic field, Parts I and II." Proc. Phys. Soc. 45 (1933) and 46 (1934).

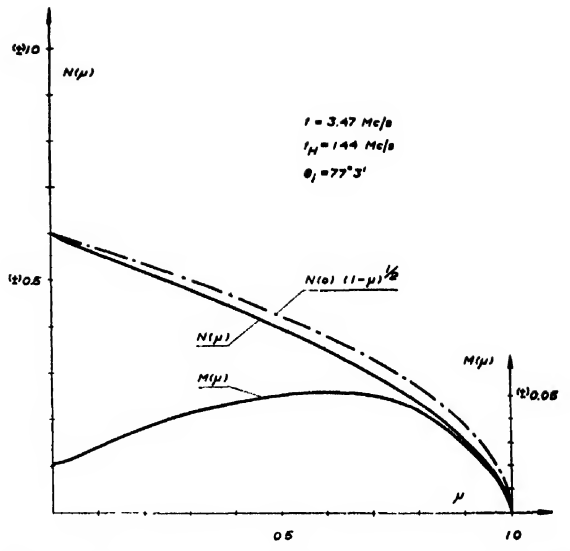


FIG. 7. Depicting the variation of $N(\mu)$ and $M(\mu)$ along the axis of the ψ-poles. $[\mu=-(\nu^2-j\delta)/j\delta_c.]$

Thus according to both the coupling theory and the geometrical interpretation one approximately has

$$\left.\begin{aligned} R_z\cong&\exp[j\cdot 2\{W_1(z_{a0}, z_0-\Delta h) \\ &+W_2(z_b, z_{a0})-\pi/4\}\cdot(T_{0z})^2, \quad \nu\leq\nu_c \\ \text{and} \\ R_z\cong&\exp[j\cdot 2\{W_1(z_b, z_0-\Delta h)-\pi/4\}]. \quad \nu\geq\nu_c \end{aligned}\right\}. \quad (29)$$

When $\nu>\nu_c$, R_0 rapidly becomes smaller and only the z and x components are recorded (if ν_c is sufficiently small).

Finally it should be mentioned that a detailed study of the connection relation between the wave functions in the n_1 and n_2 planes shows that theoretically an infinite (but highly attenuated) group of waves are running up and down between the principal branch points.[12]

However, one should not be led to believe that transmission conditions as a whole are optimal for the z wave when $\nu\cong\nu_c$. The regular path attenuation naturally becomes quite considerable unless ν_c is fairly low as at high latitudes.

The ratio between the two reflection coefficients becomes ($\nu\leq\nu_c$)

$$R_z/R_0\cong\exp[j\cdot 2\{W_2(z_{b0}, z_{a0})+W_2(z_b, z_{b0}) -W_1(z_a, z_{a0})\}]. \quad (T_{0z})^2. \quad (30)$$

Assuming $W_2(z_b, z_{b0})$ roughly equal to $W_1(z_a, z_{a0})$, a reasonable assumption when $\nu<\nu_c$, we infer from (30) that the probability of observing both the o wave and the z wave must increase when we approach such regions where ν is small and $|T|^2$ large. Let us therefore next study how $|T|^2$ varies with the collisional frequency.

As

$$T_{0z}\cong\exp\left[j\left\{k_0\int_{z_{a0}}^{z_2}(n_2-n_1)dz-(\pi/2)\right\}\right]. \quad \nu<\nu_c. \quad (27a)$$

It is of particular interest to plot n_2-n_1 from z_2 to z_{a0} (and then to z_1) along the axis between the ψ-poles. To that end we introduce

$$-\nu^2+j\delta=\mp j\mu\delta_c,$$

where the upper sign denotes the direction from z_a to z_1 and μ runs from 0 to 1. We then obtain the two simple expressions

$$n_1^2=\frac{1-(1-\mu^2)^{\frac{1}{2}}+\frac{1}{2}\mu^2\cdot\tan^2\theta}{1-(1-\mu^2)^{\frac{1}{2}}-[j\mu(1+j\delta)/y\cos\theta]}, \quad (31)$$

$$n_2^2=\frac{1+(1-\mu^2)^{\frac{1}{2}}+\frac{1}{2}\mu^2\cdot\tan^2\theta}{1+(1-\mu^2)^{\frac{1}{2}}-[j\mu(1+j\delta)/y\cos\theta]} \quad (32)$$

where it is no serious neglect to drop $j\delta$. It is therefore a comparatively simple matter to plot the variation of n_2 and n_1 along the axis between the poles. The result is shown in Fig 6. We immediately note that the difference between $\mathrm{Im}(n_2)$ and $\mathrm{Im}(n_1)$ always is very small indicating a practically negligible phase change.

Equation (27a) can be written

$$T_{0z}\cong\exp\left[-j\left\{k_0\delta_c(1-\mu)M(\mu)/\gamma_{a0}-\frac{\pi}{2}\right\}-k_0\delta_c(1-\mu)N(\mu)/\gamma_{a0}\right], \quad (33)$$

where $\mu=\nu/\nu_c$,

$$N(\mu)=\frac{1}{1-\mu}\int_\mu^1\{\mathrm{Re}(n_2)-\mathrm{Re}(n_1)\}d\mu \quad (34)$$

and

$$M(\mu)=\frac{1}{1-\mu}\int_\mu^1\{\mathrm{Im}(n_2)-\mathrm{Im}(n_1)\}d\mu. \quad (35)$$

We also have with very good accuracy

$$T_{0z}^2\cong\exp[-2k_0\delta_c(1-\mu)|N(\mu)|/\gamma_{a0}], \quad (36)$$

as at least for $|\mu|<1$, $|M(\mu)|<0.05$ (for the y and θ-values chosen). This is demonstrated by Fig. 7. The magnitude of $M(\mu)$ always is very small which indicates negligible net phase shift.

It is very interesting to note that $N(\mu)$, for the geomagnetic values chosen, is represented quite well by $N(0)(1-\mu)^{\frac{1}{2}}$. From (36) we finally obtain the practically useful formula

$$T_{0z}^2\cong\exp\left[-\frac{\nu_c\cdot\Delta h}{c_0}\left(1-\frac{\nu}{\nu_c}\right)^{\frac{1}{2}}|N(0)|\cdot\frac{\omega/\omega_{cm}}{(\omega_{cm}^2/\omega^2-1)^{\frac{1}{2}}}\right]. \quad \nu\leq\nu_c. \quad (37)$$

$$(\omega<\omega_{cm})$$

Figure 8 illustrates this result for typical E layer values and Kiruna geomagnetic conditions.

The critical collisional frequency, ν_c, increases rapidly with decreasing geomagnetic latitude as shown in Fig. 9. For comparison a height scale corresponding to the daylight ν-values has also been introduced in order to show the relative position of the ionized layer.

It is thus clear from what has been said and, again, from the nature of relation (37) that $o-z-o$ E trans-

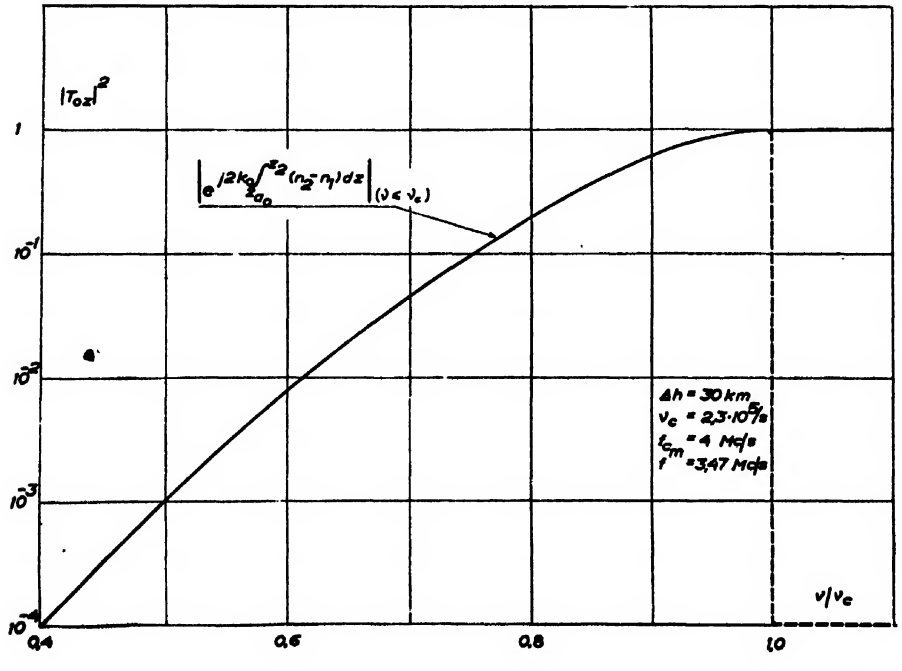


FIG. 8. The total transmission coefficient as a function of ν.

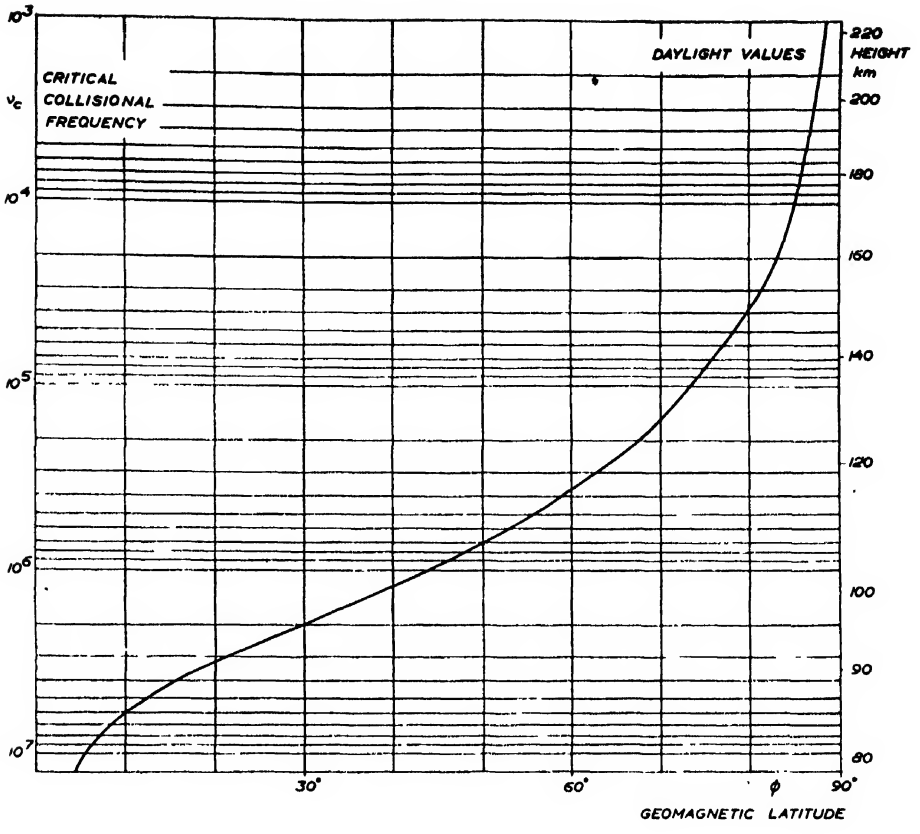


FIG. 9. The variation of critical collisional frequency with geomagnetic latitude.

mission of appreciable strength only is possible at high latitudes, preferably above 70°N or below 70°S. This is in full accord with experimental facts, both others' and our own.

EXPERIMENTAL RESULTS

It appears from Fig. 9 that at Kiruna, for example,[8] where the geomagnetic latitude is about 65°, triple splitting cannot occur with great strength much below about 120 km but that it must occur with considerable intensity slightly above that height. This is in complete agreement with our experience. We record triple penetration of the E layer almost regularly and in Fig. 10 is shown one of the typical records obtained with the new 16-kw panoramic recorder.[14] It is interesting to note that on this occasion, as so many times before, the $o-z-o$ component is almost as strong as the o component. One notes the characteristic fact that $o-z-o$ wave reflected from the F_2 layer disappears when the o wave penetrates the E layer. The reason for this is clear, both physically and from (37). This feature of the z trace has already been reported by Meek.[4] Actually the z trace is much delayed and absorbed when the o wave begins to penetrate.

It is clear both from the form of relation (37) and from Fig. 9 that at Kiruna, for example, T^2 is considerably smaller for the $o-z-o$ wave in the F_2 layer. However, on account of the small ν and therefore low attenuation on the whole, the z trace can be seen on F_2 records although usually not as frequently as near the lower critical ν_c-level. Figure 11 is a clear-cut sample of F_2 triple splitting recorded recently at Kiruna.

Figure 12 finally shows another characteristic example of E layer triple splitting which again shows that the z trace disappears near f_{E0}.

The present investigation has shown that with respect to the o and z waves when $\nu < \nu_c$ the layers of the iono-

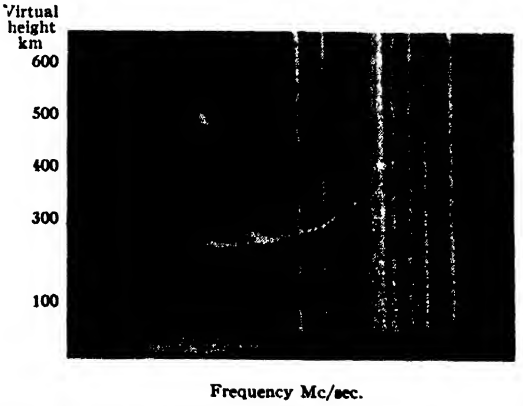


FIG. 10. Panoramic Kiruna recording of ionospheric magnetic ionic triple splitting at the E level. (April 3, 1949; 1543 GMT.)

[14] R. Lindquist, "A 16-kw panoramic ionospheric recorder," Trans. Chalmers University of Technology (in press).

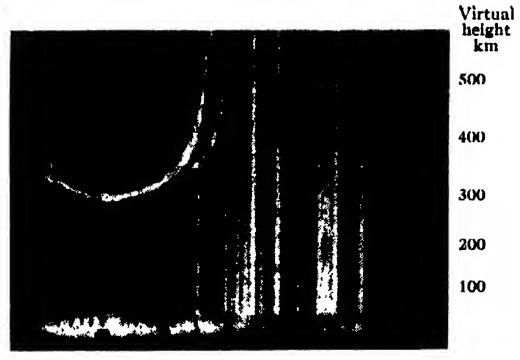


FIG. 11. Magneto-ionic F_2 triple splitting recorded at Kiruna. (October 6, 1949; 0442 GMT.)

FIG. 12. Further example of E magneto-ionic triple splitting. (April 25, 1949; 1640 GMT.)

sphere act as if they constituted coupled transmission lines. The "ordinary" line runs from $z_0 \pm \Delta h$ to $\pm z_{a0}$ the "z" line from $\pm z_{a0}$ to $\pm z_{b0}$, or from $+z_{a0}$ to $-z_{a0}$ when the $o-z-o$ wave penetrates. It is not possible to detect any difference in polarization between the o and $o-z-o$ waves outside the layers. There they are, in fact, both ordinary waves.

ACKNOWLEDGMENTS

The support of the National Research Boards for the Technical and Natural Sciences made this investigation possible. The author's thanks are also due, and are cordially extended, to S. Forsgren, E.E., who has been responsible for the preparation of the various graphs presented in this communication.

The Equivalent Circuit for the Transmission of Plane Elastic Waves through a Plate at Oblique Incidence

WILLIAM T. THOMSON
University of Wisconsin, Madison, Wisconsin
(Received April 20, 1950)

Equations are developed for the transmission of a plane elastic wave through a plate at oblique incidence with unequal fluid medium on each side of the plate. The plate transmitting both the dilatation and shear wave is reduced to a simple equivalent circuit with impedances which are functions of the incidence angle.

INTRODUCTION

THE transmission of plane elastic waves through a stratified solid medium at oblique incidence was recently treated by the author.[1] When the solid medium is reduced to a single plate surrounded by a fluid medium, the problem is somewhat simplified. The latter case has been studied by Reissner,[2] and Smyth and Lindsay.[3] Reissner's equation for transmission is conveniently expressed in terms of two factors M and N which are also used in this paper.

To make the problem more general, the fluid on each side of the plate will be assumed to be different. This, however, has no effect on the equivalent circuit, other than that of altering the terminal conditions.

Nomenclature

The symbols used are identical to those of reference 1 with few additions.

Δ', Δ'' = progressive and reflected dilatation in solid.
ω', ω'' = progressive and reflected rotations in solid.
c_Δ = velocity of dilatation wave in solid.
c_ω = velocity of rotation or shear wave in solid.
c_0 = velocity of propagation in first fluid medium.
c_2 = velocity of propagation in second fluid medium.
ρ = density of plate.
ρ_0, ρ_2 = density of fluid mediums.
p = angular frequency in radians per sec.
l, m, n = direction cosines of normal to advancing wave with respect to x, y, z axis.
σ = normal stress perpendicular to xy plane.
τ = shearing stress in xy plane.
ξ = particle velocity in z direction.
d = thickness of plate.
λ_Δ = wave-length of dilatational wave in plate.
$P=(pn_\Delta d)/c_\Delta$.
$Q=\dfrac{pn_\omega d}{c_\omega}$.
φ', φ'', φ''' = incident, reflected, and transmitted wave amplitudes in fluid.
θ_0 = angle of incidence.

$$F=\left(\frac{\cos\theta_0}{\rho_0 c_0}\right)\left(\frac{\rho c_\Delta}{\cos\theta_\Delta}\right)\left(\frac{\cos^2 2\theta_\omega}{\sin P}\right)$$

$$G=\left(\frac{\cos\theta_0}{\rho_0 c_0}\right)\left(\frac{\rho c_\omega}{\cos\theta_\omega}\right)\left(\frac{\sin^2 2\theta_\omega}{\sin Q}\right)$$

Same coefficients as those of Reissner's paper, reference 2, but rearranged in symmetric form.

$N=F+G$.
$M=F\cos P+G\cos Q$.

[1] W. T. Thomson, J. App. Phys. **21**, 89–93 (1950).
[2] H. Reissner, Helv. Phys. Acta. **11**, 140–155 (1938).
[3] J. B. Smyth and R. B. Lindsay, J. Acous. Soc. Am. **16**, 20–25 (1944).

DEVELOPMENT OF EQUATIONS

Choosing the coordinate axes as shown in Fig. 1, the equations of reference 1 apply. From Eqs. (7), (8), and (16) of Thomson's article[1] we can write the following equations corresponding to the six boundary conditions.

The shear stress at the plate boundaries is zero:

$$\tfrac{1}{2}(\Delta'-\Delta'')\sin 2\theta_\Delta+(\omega'+\omega'')\cos 2\theta_\omega=0, \tag{1}$$

$$(i/2)(\Delta'+\Delta'')\sin 2\theta_\Delta\sin P+\tfrac{1}{2}(\Delta'-\Delta'')\sin 2\theta_\Delta\cos P +i(\omega'-\omega'')\cos 2\theta_\omega\sin Q +(\omega'+\omega'')\cos 2\theta_\omega\cos Q=0. \tag{2}$$

The particle velocity is continuous at plate boundaries:

$$c_\Delta n_\Delta(\Delta'-\Delta'')-2c_\omega l_\omega(\omega'+\omega'')=c_0 n_0(\varphi'-\varphi''), \tag{3}$$

$$ic_\Delta n_\Delta(\Delta'+\Delta'')\sin P+c_\Delta n_\Delta(\Delta'-\Delta'')\cos P -2ic_\omega l_\omega(\omega'-\omega'')\sin Q -2c_\omega l_\omega(\omega'+\omega'')\cos Q=c_2 n_2\varphi'''. \tag{4}$$

The normal stress is continuous at plate boundaries:

$$\rho c_\Delta^2(\Delta'+\Delta'')\cos 2\theta_\omega-2\rho c_\omega^2(\omega'-\omega'')\sin 2\theta_\omega =-c_0^2\rho_0(\varphi'+\varphi''), \tag{5}$$

$$\rho c_\Delta^2(\Delta'+\Delta'')\cos 2\theta_\omega\cos P+i\rho c_\Delta^2(\Delta'-\Delta'')\cos 2\theta_\omega\sin P -2\rho c_\omega^2(\omega'-\omega'')\sin 2\theta_\omega\cos Q -2i\rho c_\omega^2(\omega'+\omega'')\sin 2\theta_\omega\sin Q=-c_2^2\rho_2\varphi'''. \tag{6}$$

From Eqs. (1), (2), (3), and (4) we obtain the following:

$$\omega'+\omega''=-\tfrac{1}{2}(\varphi'-\varphi'')\sin 2\theta_0, \tag{7}$$

$$\omega'-\omega''=-\frac{i}{2}(\varphi'-\varphi'')\sin 2\theta_0\frac{\cos Q}{\sin Q}+\frac{i}{2}\varphi'''\frac{c_2 n_2}{c_0 n_0}\frac{\sin 2\theta_0}{\sin Q}, \tag{8}$$

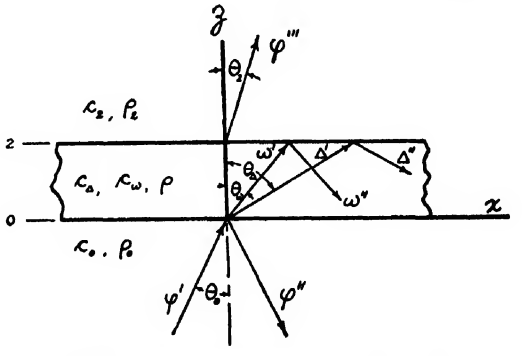


FIG. 1. Dilatational and shear waves in solid plate.

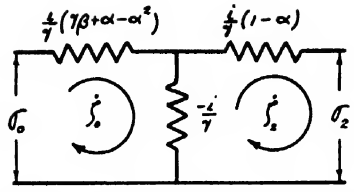


FIG. 2. Equivalent circuit for a plate transmitting dilatational and shear waves.

$$\Delta'-\Delta''=(\varphi'-\varphi'')\sin2\theta_0\frac{\cos2\theta_\omega}{\sin2\theta_\Delta}, \tag{9}$$

$$\Delta'+\Delta''=i(\varphi'-\varphi'')\frac{\sin2\theta_0\cos2\theta_\omega\cos P}{\sin2\theta_\Delta\sin P}-i\varphi'''\frac{c_2n_2}{c_0n_0}\frac{\sin2\theta_0\cos2\theta_\omega}{\sin2\theta_\Delta\sin P}. \tag{10}$$

Substituting these into Eqs. (5) and (6) we obtain

$$\varphi'-\varphi''=\frac{\varphi'''}{N}\frac{c_2}{c_0}\left[M\frac{n_2}{n_0}+i\frac{c_2\rho_2}{c_0\rho_0}\right], \tag{11}$$

$$\varphi'+\varphi''=\frac{\varphi'''}{N}\frac{c_2}{c_0}\left[M\frac{c_2\rho_2}{c_0\rho_0}+i(N^2-M^2)\frac{n_2}{n_0}\right], \tag{12}$$

$$\frac{\varphi'''}{\varphi'}=\frac{2N(c_0/c_2)}{M\left(\frac{c_2\rho_2}{c_0\rho_0}+\frac{n_2}{n_0}\right)+i\left[(N^2-M^2)\frac{n_2}{n_0}+\frac{c_2\rho_2}{c_0\rho_0}\right]}. \tag{13}$$

$$T=\frac{c_2\rho_2|\varphi'''|^2}{c_0\rho_0|\varphi'|^2}=\frac{4N^2(c_0/c_2)(\rho_2/\rho_0)}{M^2\left(\frac{n_2}{n_0}+\frac{c_2\rho_2}{c_0\rho_0}\right)^2+\left[\frac{c_2\rho_2}{c_0\rho_0}+(N^2-M^2)\frac{n_2}{n_0}\right]^2}, \tag{14}$$

$$\frac{\varphi''}{\varphi'}=\frac{M\left(\frac{c_2\rho_2}{c_0\rho_0}-\frac{n_2}{n_0}\right)+i\left[(N^2-M^2)\frac{n_2}{n_0}-\frac{c_2\rho_2}{c_0\rho_0}\right]}{M\left(\frac{c_2\rho_2}{c_0\rho_0}+\frac{n_2}{n_0}\right)+i\left[(N^2-M^2)\frac{n_2}{n_0}+\frac{c_2\rho_2}{c_0\rho_0}\right]}, \tag{15}$$

$$R=\frac{c_2\rho_2|\varphi''|^2}{c_0\rho_0|\varphi'|^2}=\left\{\frac{M^2\left(\frac{c_2\rho_2}{c_0\rho_0}-\frac{n_2}{n_0}\right)^2+\left[(N^2-M^2)\frac{n_2}{n_0}-\frac{c_2\rho_2}{c_0\rho_0}\right]^2}{M^2\left(\frac{c_2\rho_2}{c_0\rho_0}+\frac{n_2}{n_0}\right)^2+\left[(N^2-M^2)\frac{n_2}{n_0}+\frac{c_2\rho_2}{c_0\rho_0}\right]^2}\right\}\frac{c_2\rho_2}{c_0\rho_0}. \tag{16}$$

Equations (14) and (16) are the transmission and reflection coefficients of the plate. When the two fluid mediums are the same, these equations reduce to

$$T=\frac{|\varphi'''|^2}{|\varphi'|^2}=\frac{4N^2}{4M^2+(N^2-M^2+1)^2}, \tag{17}$$

$$R=\frac{|\varphi''|^2}{|\varphi'|^2}=\frac{(N^2-M^2-1)^2}{4M^2+(N^2-M^2+1)^2}, \tag{18}$$

which agree with those of Reissner.

DEVELOPMENT OF THE EQUIVALENT CIRCUIT

In the impedance theory of tubes and horns[4] it is common practice to write the pressure and volume current at one point in terms of corresponding quantities at another point. This idea was utilized in the paper of reference 3 to consolidate the transmission equations. With this same general scheme, it is possible to go further and develop the equivalent circuit for the plate transmitting the dilatation and shear waves.

We first write the equations for σ and $\dot{\xi}$ at the plate boundaries in terms of φ', φ'', and φ''' by substituting Eqs. (7), (8), (9), and (10) in the left side of Eqs. (3), (4), (5), and (6).

$$\sigma_0=i(\varphi'-\varphi'')\rho_0c_0^2M-i\varphi'''\rho_0c_0^2N(c_2n_2/c_0n_0), \tag{19}$$

$$\dot{\xi}_0=(\varphi'-\varphi'')c_0\cos\theta_0, \tag{20}$$

$$\sigma_2=i(\varphi'-\varphi'')\rho_0c_0^2N-i\varphi'''\rho_0c_0^2M(c_2n_2/c_0n_0), \tag{21}$$

$$\dot{\xi}_2=\varphi'''(c_2n_2/n_0)\cos\theta_0. \tag{22}$$

[4] Stewart and Lindsay, *Acoustics* (D. Van Nostrand Company, Inc., New York, 1930), 4th edition, Appendix III and IV.

We next seek a solution in the form

$$\sigma_2 = \alpha\sigma_0 - i\beta\dot{\zeta}_0, \tag{23}$$

$$\dot{\zeta}_2 = \alpha\dot{\zeta}_0 - i\gamma\sigma_0. \tag{24}$$

Substituting Eqs. (19) to (22) in Eqs. (23) and (24) and equating coefficients of like terms of $(\varphi' - \varphi'')$ and φ''' we obtain the following

$$\alpha = M/N, \tag{25}$$

$$\beta = (\rho_0 c_0/\cos\theta_0)(M^2 - N^2)/N, \tag{26}$$

$$\gamma = -(\cos\theta_0/\rho_0 c_0)1/N, \tag{27}$$

$$\alpha^2 + \beta\gamma = 1. \tag{28}$$

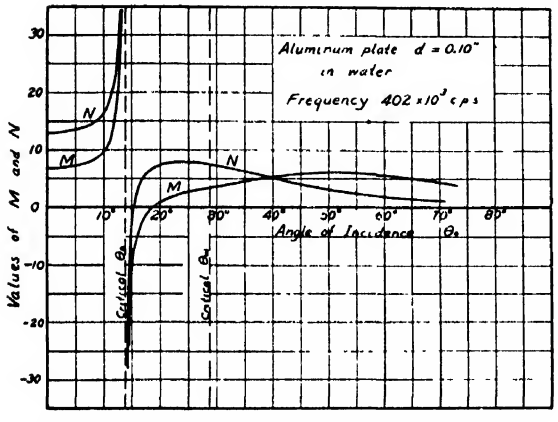


FIG. 3. Values of M and N for aluminum plate in water.

It is thus possible to write the circuit equations

$$\sigma_0 - \sigma_2 = (i/\gamma)(1-\alpha)\dot{\zeta}_2 + (i/\gamma)(\gamma\beta + \alpha - \alpha^2)\dot{\zeta}_0, \tag{29}$$

$$\sigma_2 = (-i/\gamma)(1-\alpha)\dot{\zeta}_2 + (-i/\gamma)(\dot{\zeta}_0 - \dot{\zeta}_2), \tag{30}$$

from which the equivalent circuit of Fig. 2 is obtained.

For this problem, the impedance of the fluid medium 2 is $\sigma_2/\dot{\zeta}_2 = -(\rho_2 c_2)/\cos\theta_2$ while that of medium 0 is $\sigma_0/\dot{\zeta}_0 = -(\rho_0 c_0/\cos\theta_0)(\varphi' + \varphi''/\varphi' - \varphi'')$. The circuit impedances are only functions of N, M, and the quantity $(\rho_0 c_0)/\cos\theta_0$.

Computation

The equations show that M and N contain only one singularity at the critical angle of the dilatational wave. The transmission at this angle is, however, not zero.

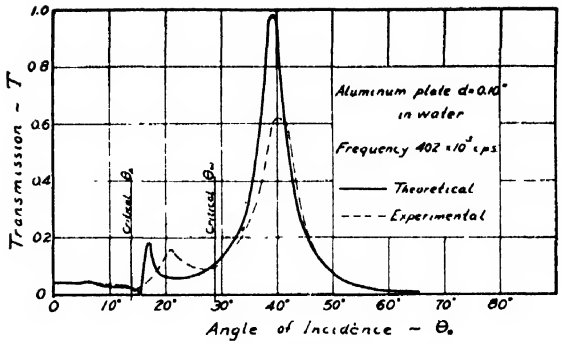


FIG. 4. Transmission curves for aluminum plate in water.

Also, for small ratios of the plate thickness to the wavelength, M and N are nearly equal for all values of θ_0, and further simplification of the transmission equations and the equivalent circuit is possible. The transmission becomes equal to zero only when $N = 0$.

Due to the rather lengthy computations involved, theoretical calculations are shown only for a single case of an aluminum plate 0.10-inch thick immersed in water with an ultrasonic beam of 402 kilocycles/sec. The thickness ratio for the plate at this frequency is $d/\lambda_\Delta = 0.1628$, which cannot be classified in the thin plate region where $N = M$. For an aluminum plate $d = 0.015$ inch at the same frequency, the curves for N and M practically coincide.

The curves for N and M shown in Fig. 3 are typical in that the singularity is evident at the critical angle of the dilatation wave and the point $N = 0$ corresponding to zero transmission occurs at a slightly larger angle. The corresponding transmission curve is shown in Fig. 4 with a comparison with experimental values taken from the dissertation* of Mr. Gerald Cohen. The transmission peak occurs near 40°, which is well beyond the critical angles for dilatation and shear. The disagreement in the peak values is evidently due to dissipation which is not accounted for in the theoretical equations.

The author is indebted to the staff of the Computing Service of the University of Wisconsin for the calculations and to the Navy Research Laboratory of the University for sponsoring the investigation carried out by Mr. Cohen.

* Gerald H. Cohen, "Ultrasonic Transmission through Thin Plates," Ph.D. thesis, University of Wisconsin Engineering College, 1950.

Gas Evolution in Liquids and Cavitation

P. H. SCHWEITZER*
The Pennsylvania State College, State College, Pennsylvania
AND
V. G. SZEBEHELY†
Virginia Polytechnic Institute, Blacksburg, Virginia
(Received February 8, 1950)

Nine heavy lubricating oils, four light lubricating oils, three aircraft engine fuels, one diesel fuel, and distilled water were tested for their air solubility and for the rate of solution and evolution of air when equilibrium is disturbed. The technique consisted of shaking of a container partly filled with liquid and taking accurate pressure measurements of the air above the liquid. Since air in solution exerts no pressure, the variation of pressure during and after the agitation furnished information on the amount of air that went into or came out of solution and their time rates. The solution was considered to have reached equilibrium when further shaking no longer changed the air content of the liquid.

A wide variation exists in the solubility constants and rates, but the rate of evolution was always found to be proportional to the supersaturation, and the rate of solution to the undersaturation, under otherwise similar conditions. At room temperature the air solubility constant for distilled water is 1.84 percent, for a gasoline approximately 20 percent, and for (light and heavy) lubricating oils approximately 10 percent. The "half-life" for evolution on the other hand was 3.86 sec. for distilled water, 0.13–0.24 sec. for gasoline, 3.6–7.6 sec. for light, and 11.6–51.4 sec. for heavy lubricating oils.

The investigation gave some insight into the phenomenon of cavitation which is still largely shrouded in mystery.

SYMBOLS

t = Time (sec.)
p_0 = Atmospheric pressure (p.s.i.)
p = Gas pressure above the liquid generally (at time t) (p.s.i.)
p_e = Gas pressure above the liquid at the beginning of the evolution or solution process (p.s.i.)
p_e, p_e' = Equilibrium or saturation pressure (gas pressure above the liquid in case of equilibrium) (p.s.i.)
Δp = Pressure difference: $p_e - p_0$ (p.s.i.)
V_l = Volume of used liquid (cu. ft.)
V = Volume of free gas dissolved in the liquid generally (at time t) (cu. ft.)
V_g = Volume of the gas section of the container (cu. ft.)
V_e, V_e' = Volume of free gas dissolved in the liquid at equilibrium (cu. ft.)
ΔV = Volume of free gas evolved from the liquid in time t (cu. ft.)
S = Solubility constant —
r = Gas-liquid volume ratio —
R_E = Rate of evolution (1/sec.)
R_S = Rate of solution (1/sec.)
T_E = Half-life of evolution (sec.)
$(T_E)_0$ = Limiting value of half-life, if $r \to \infty$ (sec.)
T_S = Half-life of solution (sec.)
s, $(-s)$ = Supersaturation (undersaturation) —.

THREE possible relations can exist between the saturation pressure and the gas pressure above the liquid:

(a) $p_e = p_0$. Case of equilibrium. The liquid can neither release nor absorb more gas.

(b) $p_e > p_0$. Case of supersaturation. The amount of gas dissolved in the liquid is greater than the amount which would correspond to the existing gas pressure. Therefore, evolution process *can* take place.

(c) $p_e < p_0$. Case of undersaturation. The amount of gas dissolved in the liquid is less than the amount which would correspond to the existing gas pressure. Therefore, solution process *can* take place. (It will be discussed later that evolution or solution process does not necessarily take place and the existence of that depends upon several other factors.)

The case of equilibrium obeys the well-known Henry's (Henry-Dalton's) law which says that if no change in molecular structure occurs during the solution or evolution process, then the free volume of dissolved gas in the liquid is directly proportional to the absolute pressure

$$V_e = S \frac{p_e}{p_0} V_l. \tag{1}$$

Next the case of supersaturation is considered. A container is partly filled with saturated liquid and gas, the gas pressure being equal to the saturation pressure. If the gas pressure is slightly reduced, diffusion and surface evasion will take place, which will be called "quiescent evolution." No bubble formation will accompany this evolution process. Our experiments have shown that the ratio between the rates of evolution obtained under agitated and quiescent conditions is always in the thousands, therefore in the following discussion, the "quiescent evolution" will be neglected.

Considering the slowly obtained small pressure drop and neglecting the quiescent evolution, a slightly supersaturated liquid is obtained. If the container is kept at rest and the pressure is further reduced very slowly to atmospheric pressure, then a supersaturated liquid is obtained. We wish to point out that considerable (up to 100 percent) supersaturation was often produced without any bubble formation.

If the pressure drop is produced suddenly, rapid evolution does take place. The liquid surface suffers a break which is equivalent to agitation and immediate bubble formation can be observed. If the liquid is well stirred or agitated, supersaturated liquid cannot be produced. If we consider breaks in the liquid surface, caused by a sudden pressure drop, as a special sort of

* Professor of Engineering Research.
† Associate Professor of Applied Mechanics.

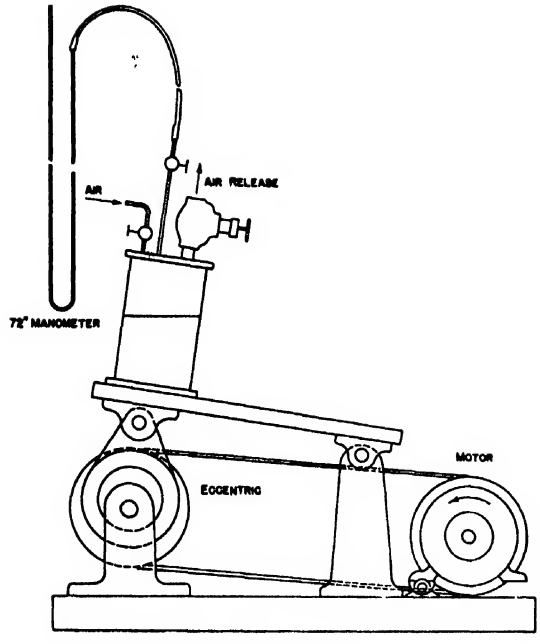


FIG. 1. Shaker.

agitation, it may be stated that bubble formation in a moderately supersaturated liquid can only set in if the liquid is agitated.

RATE OF EVOLUTION

The rate of evolution is defined as the time rate of change of the volume of free gas in solution per unit volume of the liquid

$$R_E=\frac{d}{dt}\left(\frac{V}{V_l}\right). \tag{2}$$

In order to obtain the $V(t)$ function, the following experiment was done:

First the p_e equilibrium or saturation pressure was slowly reduced to p_g, keeping the container in rest and so avoiding any bubble formation. Then agitation was started, such as shaking of the container as shown in Fig. 1. A cylindrical container of 3400 cc was shaken with a frequency of 400 cycles per minute and with a stroke of one inch. With the agitation, the evolution process starts and gas leaves the liquid. To measure the volume of the evolved gas, two methods have been used.

In the direct method, the gas pressure before the evolution process would start, must be brought to atmospheric: $p_g=p_0$. The container then is closed and connected with a volume measuring instrument. The amount of evolved air is recorded as a function of the time, keeping the pressure above the liquid always atmospheric. The original volume of dissolved air (before agitation) can be computed from formula (1). If the instantaneously recorded volume of evolved free gas is V_m, then

$$V=V_e-V_m.$$

Because accurate instantaneous recording of V_m involved experimental difficulties, we used an indirect method based on pressure measurement. The gas pressure above the liquid pressure was reduced from p_e equilibrium pressure to p_g gas pressure, which in this method need not be atmospheric. Then the container was closed and connected to a manometer. The effect of agitation was that the pressure (p) increased from p_g to a new equilibrium pressure p_e'. Obviously $p_g<p_e'<p_e$, i.e., the new equilibrium pressure is greater than the reduced gas pressure, but less than the original equilibrium pressure; they are related by Eq. (3):

$$V=S\frac{p_e'}{p_0}V_l-\frac{p-p_e'}{p_0}V_g. \tag{3}$$

The p_e' final equilibrium pressure can also be predicted from the equilibrium equation as:

$$p_e'=\frac{V_g p_g+V_l S p_e}{V_g+V_l S}=\frac{r p_g+S p_e}{r+S}.$$

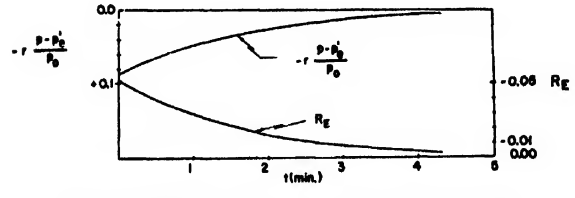


FIG. 2. Pressure *vs.* time and rate *vs.* time curves for lubricating oil.

Substituting Eq. (3) in Eq. (2), the rate of evolution is

$$R_E=-r\frac{d}{dt}\frac{p-p_e'}{p_0}.$$

Knowing r, p_e', p_0 and measuring p, the rate of evolution, R_E, can be obtained as a function of the time by graphical differentiation.

Such a plot, for a lubricating oil, is shown in Fig. 2. As the pressure appears to approach the saturation pressure p_e', the rate of evolution R_E is approaching zero.

It did not seem unreasonable to assume that the rate of evolution is always proportional to the supersaturation (and the rate of solution to undersaturation):

$$R_E=kS=k\frac{V_e'-V}{V_l}. \tag{4}$$

If this were the case Eq. (2) could be integrated with the result:

$$\frac{d}{dt}\left(\frac{V}{V_l}\right)=k\frac{V_e'-V}{V_l}$$

$$V=(V_e-V_e')e^{-kt}+V_e'. \tag{5}$$

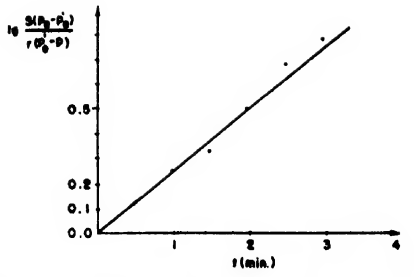


FIG. 3. Pressure *vs.* time curve plotted on semilog base.

If assumption (4) is correct then plotting $\lg(V-V_e')$ against time t should give a straight line. After a simple transformation we get in place of Eq. (5)

$$-\frac{p-p_e'}{p_0}=-\frac{S}{r}\frac{p_e-p_e'}{p_0}e^{-kt}, \tag{6}$$

therefore plotting $\lg[S(p_e-p_e')/r(p_e'-p)]$ against t should also give a straight line which should go through the origin.

Figure 3 shows a plot for the above mentioned lubricating oil which was selected at random. In the course of the tests performed on 18 liquids some 150 of such plots were made for both solution and evolution runs. Except for slight inaccuracies, attributable to experimental errors, all turned out to be straight lines going through the origin. The conclusion was clear: the rate of evolution *is* proportional to the supersaturation (the rate of solution to the undersaturation) and represents, therefore, an exponential function of time.

Inspecting Eq. (5) it is seen that at zero time $V=V_e$ and if $t\to\infty$, $V\to V_e'$. Finally the total evolved gas is V_e-V_e' but infinite time is required for its evolution. Consider the time required for the evolution of one-half of the total evolved gas $(V_e-V_e')/2$ and designate it by the symbol T_E. When $t=T_E$, $V_e-V=(V_e-V_e')/2$, therefore from Eq. (5)

$$k=\frac{ln2}{T_E}=\frac{0.693}{T_E}. \tag{7}$$

T_E is designated "half-life" of evolution and is equal to the time required to half complete the gas evolution. With this new designation the gas content of the supersaturated liquid is:

$$V=(V_e-V_e')\exp\left(-0.693\frac{t}{T_E}\right)+V_e'. \tag{8}$$

In an actual case the V_e and V_e' values are not known —at least they have to be computed using the

$$V_e=S\frac{p_e}{p_0}V_l \quad \text{and} \quad V_e'=S\frac{p_e'}{p_0}V_l$$

formulas. Equation (8) can be transformed to a more useful form

$$V=\frac{S}{r+S}\frac{V_l}{p_0}\left[r(p_e-p_0)\exp\left(-0.693\frac{t}{T_E}\right)+rp_0+Sp_e\right]. \tag{9}$$

With formula (9), V, the value of free gas still dissolved in the liquid at time t can be evaluated. In many engineering applications, however, it is more important to know the amount of the evolved gas from $t=0$ to $t=t$; $\Delta V=V_e-V$. This is given by:

$$\Delta V=\frac{Sr}{S+r}\frac{p_e-p_0}{p_0}V_l\left[1-\exp\left(-0.693\frac{t}{T_E}\right)\right]. \tag{10}$$

Formula (10) furnishes two as yet unmentioned results. The $Sr/(S+r)$ factor can be represented by hyperbolas going through the origin, whether S or r is considered as the variable.

Due to the fact that the asymptotes are parallel with the horizontal axis, the effect of change in r (keeping S constant) is appreciable if r is small, but it can be neglected if r is great, (Figure 4.). For thin liquid layers $(r\to\infty)$ the $Sr/(S+r)$ value is represented by S. The last factor, $1-\exp(-0.693\,t/T_E)$ is the well-known exponential function starting with zero value at time zero and approaching to 1 if $t\to\infty$.

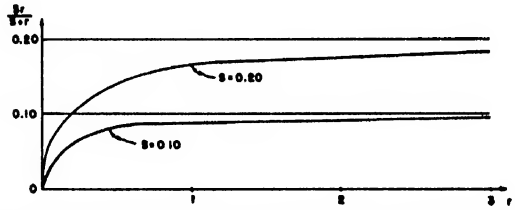


FIG. 4. The $Sr/(S+r)=f(r)$ hyperbolas for different S values.

The initial tangent (at $t=0$) to the curve is $0.693/T_E$ which together with Fig. 5 show that if the half-life time is great the evolution is slow.

An example with actually obtained values explains better the $V=f(t)$ function.

For the heavy lubricating oil referred to, the solubility constant S was 0.0815 with respect to air. The used air-oil volume ratio was $r=2$. The measured half-life was $T_E=1.205$ min. The volume of used liquid was $V_l=63.5$ cu. in. The initial saturation pressure was $p_e=38.45$ p.s.i.g. and the air pressure above the oil $p_0=21.86$ p.s.i.g.

With the above figures from formula (10):

$$\Delta V=5.6[1-\exp(-0.575t)]$$

if ΔV is in cu. in. and t in minutes. Formula (9) gives the free volume of air in solution. Substituting the numerical values: $V=5.6\exp(-0.575t)+13.1$.

Figure 6 shows the volume of evolved air and Fig. 7 the volume of air in solution in function of time.

Going back to the definition of the rate of evolution and substituting in formula (2) for V the value obtained in formula (9), it is found that the rate is a function of the half-life (T_E), of the solubility constant (S), of the gas-liquid ratio (r), of the pressure difference ($p_e - p_0$), of the atmospheric pressure (p_0) and of the time (t). It was investigated experimentally whether these variables are related to each other or are independent. It was found that the half-life is related to the gas-liquid ratio and the solubility constant but none of the other variables involved are related to each other.

For a given liquid and gas combination S is the only constant with which the evolution process can be described inasmuch as S is not a function of the other variables. Even the solubility constant varies with temperature, but in this investigation the temperature was kept constant. For the same gas-liquid combination, the half-life varies with the gas-liquid volume ratio. Therefore half-lives can be compared only at a certain r

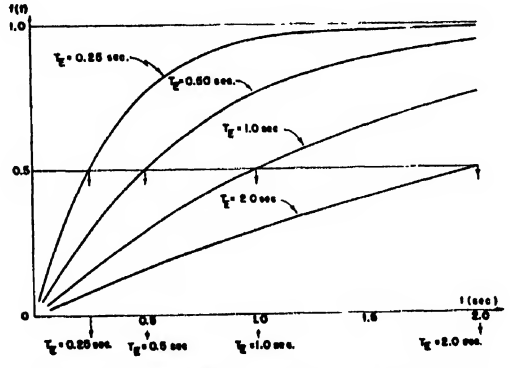


FIG. 5. The $1-e^{-0.693(t/T_E)}=f(t)$ exponential curves for different T_E values.

value. Obviously the half-life depends also upon the method of agitation, therefore a standardized agitation is required in order to compare data. It was found that the relation between T_E and r can be described with an exponential function. Increasing the gas-liquid volume ratio by varying r from very small values up to 4.5 the following relation was indicated:

$$T_E=(T_E)_0/(1-e^{-cr}),$$

where $(T_E)_0$ is limiting value of half-life if $r\to\infty$, i.e., if we deal with a very thin liquid film, and c is a constant. On the other hand, keeping r constant, a relation is expected between the half-life, solubility constant, viscosity, density, surface tension, etc. These investigations are extended to other details too and at the present time are conducted at the Virginia Polytechnic Institute, Department of Applied Mechanics.

SOLUTION

After the foregoing detailed discussion of the evolution process, the solution process can be settled very shortly. In the solution process a disturbed undersaturated liquid is approaching to an equilibrium condition. The analogy between evolution and solution processes is complete, therefore the same letters and steps can be used in deriving the same formula.

The gas pressure above the liquid p_g is now greater than the saturation pressure p_e. Closing this p_g pressure in the container and connecting it with a manometer or pressure gauge, after an agitation, the p_g pressure will drop to a p_e' new equilibrium pressure and some of the gas originally above the liquid will be dissolved. The definition of the rate of solution is $R_S=d/dt(V/V_l)$. The rate of solution will always be positive, due to the fact that the free volume of dissolved gas is increasing if the time is increasing. During the solution process, the pressure drops from p_g to p_e' and $p_g>p_e'>p_e$.

The $T_E(r, S)$ function was discussed already and the same can be said about the $T_S(r, S)$ function. It should not be a surprise however that using the same liquid and gas (S) as well as the same gas-liquid volume ratio (r) for an evolution and a solution process, the half-lives will not be the same. The experiments performed showed that with no exception the half-life for evolution was always shorter than that for solution, which means that the evolution process is quicker than the solution process. This difference between T_S and T_E can be explained by considering the mechanism of evolution and solution. The mechanism actually changes the effective value of r during the evolution process by exposing more surface through bubble formation, but it results in only negligible changes in r during solution. A detailed analysis of this question will be published in another paper.

RESULTS

The experimental investigation included 18 liquids and one gas, air. The liquids tested were heavy and light lubricating oils, a diesel fuel, distilled water and aircraft engine fuels. The smallest value of the solubility constant (1.84 percent) was obtained for water. For heavy lubricating oils it varied between 7.75 percent and 9.18 percent, for light oils between 9.70 percent and 11.30 percent. For the tested diesel fuel S was roughly 12 percent and the maximum obtained

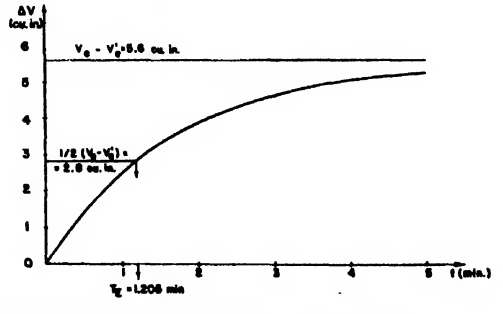


FIG. 6. Volume of evolved air for heavy lubricating oil *vs.* time. Half-life, $T_E=1.205$ min.; $S=0.0815$; $r=2$; $V_l=63.5$ cu. in.; $p_e=38.45$ p.s.i.g.; $p_g=21.86$ p.s.i.g.

value of S was 25.14 percent in case of an aircraft engine fuel.

Discussing results obtained regarding half-lives, first a standard agitation has to be established, then a certain gas-liquid ratio has to be decided at which the values can be compared. The following values of half-lives correspond to a gas-liquid ratio, $r=4$. The frequency used was 400 cycles per minute and the stroke one inch. Without any special temperature controlling equipment the results were reproducible; the usual room temperature variations did not seem to effect the results.

For heavy oils, the solution and evolution process is slow, therefore the half-lives are long. As it was explained, the evolution process is quicker than the solution. The shortest half-lives were obtained for evolution of fuels (0.128 sec.) and the longest for solution of heavy oils (414.0 sec.). In Table I the solubility constants and the half-lives for solution as well as for evolution are shown for the tested liquids.

From the tabulated data, the behavior of the liquids can be predicted after a brief and simple computation.

RELATION BETWEEN CAVITATION AND GAS EVOLUTION

In ordinary practice liquids display no tensile strength and when the absolute pressure in a liquid drops below zero, the liquid tears apart. In fact, discontinuity in the form of vapor-filled bubbles, usually occurs as soon as the pressure in the liquid is reduced to its vapor pressure. Hence, cavitation is generally associated with the vapor pressure of the liquid and is expected to take place whenever and wherever the hydraulic pressure reaches vapor pressure.

Our observations with bubble formation have convinced us that this widely held view generally is incorrect, though it appears to be valid under certain specific circumstances.

Cavitation is failure (of continuity) of the liquid in tension. Like failure in solids, it is largely controlled by contamination and surface conditions. The true tensile strength of liquids (no contamination and no surface flaw) is extraordinarily high. Our information on the tensile strength of liquids is very limited because of the practical difficulty of subjecting a column of pure liquid to tension. The great difference between solids and liquids is not so much in their tensile strengths, but in their shearing strength. To test a solid rod in tension we clamp its two ends and pull them apart. The absence of shearing strength prevents us from clamping a liquid column. A liquid cannot be handled without a container. The adhesion between the liquid and the container is frequently less than the cohesion between the liquid particles themselves. Under such circumstances, the true negative pressure which is necessary to tear liquid particles apart cannot be determined. Such tests can yield only a minimum value because in measuring the tensile strength of a chain one learns something about its weakest link, nothing more.[1]

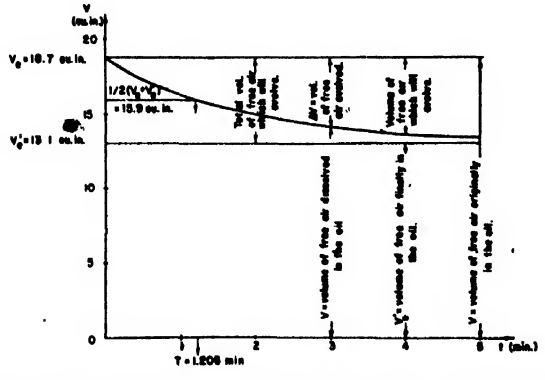


FIG. 7. Free volume of air in solution for heavy lubricating oil *vs.* time. (Specifications same as on Fig. 6.)

The weak link may be somewhere else. If the liquid is not completely degassed, it will usually contain some entrained air in the form of microscopic or submicroscopic bubbles either in the bulk of the liquid or near a solid contaminant or near the container wall. Such a bubble, whatever its size, represents on the one hand a flaw within the liquid and on the other hand, an adjacent gas phase into which dissolved gas can diffuse or liquid vapor can evaporate. Under such a condition, visible cavitation (macrocavitation) will take place easily and promptly. However, this represents a "false cavitation in which the cavities were derived from pre-existing gas nuclei rather than de novo."[2]

In spite of all these experimental handicaps it has been shown[2,3] that liquid water can stand a negative pressure of 50 atmospheres and more without visible failure.

Weyl and Marboe[1] tested the gas release from water saturated with carbon dioxide at atmospheric pressure. They subjected the CO_2 saturated water to a partial vacuum. When the liquid was held in an ordinary "clean" glass container, a pressure drop from 760 to 50 mm Hg abs. readily produced gas bubbles. However, if the glass container was thoroughly cleaned with dichromate-sulphuric acid solution, bubbles failed to appear upon evacuation. Clean glass powder in the liquid did not cause the release of gas bubbles either, but glass powder which had been exposed to laboratory air for a short time did. Also when the surface of the container was not thoroughly cleaned, release of gas supersaturation promptly accompanied the drop in pressure. This directed attention to the chemical structure of the contacting phases.

In our tests the container used was made of steel and Lucite. No precaution was made to have them chemically clean. With water we were unable to produce any appreciable supersaturation without observing bubble formation. However, with petroleum hydrocarbons,

[1] Weyl and Marboe, Research **2**, 19 (1949).
[2] Pease and Blinks, J. Phys. Colloid Chem. **51**, 556, No. 2 (1947).
[3] E. N. Harvey, J. App. Phys. **18**, 161 (1947).

which wet both steel and Lucite, we observed considerable supersaturations (100 percent) without bubble release, provided the liquid was kept in a quiescent state. However, when the liquid was agitated, gas supersaturation was released rapidly.

Gas may be present in a liquid in two forms, dissolved gas and entrained gas. The dissolved gas is actually no gas at all. It is invisible (the liquid is clear), and does not increase its volume and compressibility by any appreciable amount. Entrained gas is dispersed in the liquid in the form of bubbles which may give the liquid a turbid appearance. The equilibrium amount of entrained air (or ambient gas) in a quiescent liquid is zero. Churning tests showed that it reaches a maximum at certain degree of agitation and decreases again when that degree of agitation is exceeded.

From these observations the following picture emerges. No vapor-cavitation can take place without entrained air (or other gas) in the liquid. As soon as there is an air bubble, that bubble promptly fills up with vapor, provided the hydraulic pressure in the liquid is lower than its vapor pressure. As long as there is no air bubble, no vapor can form (except on the surface if the liquid has one). Evaporation takes place when liquid in the vapor phase leaves the liquid in the liquid phase. This requires an interface. Vapor cannot leave the liquid unless it goes some place. If there are air bubbles, it goes into the air bubbles through the liquid-air interface. Without air nuclei there is no evaporation (except on top surface if there is one) and no cavitation. With air nuclei present, bubble growth through gas release will set in as soon as the hydraulic pressure drops below the saturation pressure which corresponds to the air actually held in solution.

It has generally been observed that in flowing water cavitation bubbles appear whenever and wherever the hydraulic pressure drops below the vapor pressure of the water, approximately 0.36 p.s.i.a. at 70°F. This observation is the consequence of two extraneous circumstances. The first is that flowing water always contains some entrained air in the form of small bubbles either in the bulk of the liquid or adjacent to solids in contact with the water. The second is that in water, having a low air solubility, the release of dissolved air (rate of evolution) is very slow if the pressure is reduced a small amount from the saturation pressure. Therefore, the bubble growth is very slow. If the bubbles were submicroscopic at the start of the experiment (at saturation pressure), they will remain submicroscopic when the pressure is reduced a small fraction of an atmosphere. However, when the pressure reaches the vapor pressure, there is a relatively fast vapor release into the air bubbles and then their growths become visible to ordinary observation. The term "cavitation" is frequently applied to this phase of bubble growth. Then cavitation appears to set in at a locality where the pressure of the flowing water drops below its vapor pressure.

TABLE I. Summary of experimental results.

		Solubility constant (%)	Half-life at $r=4$ for evolution solution (seconds)		Kin. viscosity at 70°F (centistokes)
1	Heavy lubricating oil	7.75	51.4	414	823–974
2	Heavy lubricating oil	8.14	43.0	292	823–974
3	Heavy lubricating oil	8.15	45.0	386	762–974
4	Heavy lubricating oil	8.24	48.2	396	762–974
5	Heavy lubricating oil	8.61	37.2	244	762–974
6	Heavy lubricating oil	8.62	34.8	252	562–670
7	Heavy lubricating oil	9.05	31.0	125.4	325–411
8	Heavy lubricating oil	9.11	17.9	50.2	249–346
9	Heavy lubricating oil	9.18	11.6	32.7	88.7
10	Light lubricating oil	9.70	4.74	10.22	17.5
11	Light lubricating oil	9.95	4.18	9.77	17.5
12	Light lubricating oil	10.72	3.56	6.14	13.5
13	Light lubricating oil	11.30	7.63	9.46	17.5
14	Diesel fuel	11.98	0.301	3.08	4.4
15	Aircraft engine fuel	17.20	0.236	2.595	1.67
16	Aircraft engine fuel	22.80	0.128	1.214	0.6350
17	Aircraft engine fuel	25.14	0.137	1.344	0.6350
18	Distilled water	1.84	3.86	7.930	1.0

In order to accurately describe what takes place it is necessary to define our terms. A liquid is considered to be in *equilibrium* when its content of dissolved gas is constant with respect to time. *Cavitation* is any discontinuity in the bulk of the liquid, irrespective of the size of the cavity and whether it is filled with void, gas, or vapor. *Microcavitation* refers to bubbles of microscopic or submicroscopic size. *Macrocavitation* refers to visible bubbles. The liquid is *quiescent* when all of its parts are at rest relative to the container, except for their thermal motion. The liquid is in *agitation* when some or all of it is not quiescent.

With these definitions:

1. A liquid is in equilibrium if it has been either quiescent or under an invariable type of agitation for a sufficient length of time to assume saturation in all parts. The equilibrium is upset if either the pressure or the temperature changes locally or in the whole.

2. A quiescent liquid in equilibrium with air has a zero amount of entrained air and a saturation amount of dissolved air.

3. An agitated liquid in equilibrium has a certain amount of entrained air and a saturation amount of dissolved air.

4. A quiescent liquid, originally in equilibrium, will not cavitate until it is subjected to a negative pressure which is equal to the tensile strength of the liquid (not less than −750 p.s.i. for water).

5. A liquid in agitation and exposed to air for any length of time cavitates continuously inasmuch as it contains entrained air in the form of nuclei (microcavitation).

6. In the bulk of the liquid (a place other than the surface) neither gas evolution nor vapor evolution can take place in the absence of gas nuclei (microcavitation).

7. In an originally gas free liquid, microcavitation may set in (1) if air is introduced into it under the surface, (2) if the liquid is supersaturated *and* has chemically weak bonds with the container wall or solid contaminants, (3) if it is subjected to agitation.

8. Microcavitation will turn into macrocavitation as a result of either gas evolution, or vapor evolution, or both. Gas evolution begins at any pressure lower than saturation pressure; vapor evolution begins only below the vapor pressure of the liquid.

This explains the behavior of flowing water under ordinary conditions. Under some agitation the water contains a certain amount of entrained air in the form of microscopic or submicroscopic bubbles which are invisible. Ordinarily it also contains some air nuclei in the microscopic crevices of the container, submerged

solids, and solid contaminants. If the hydraulic pressure drops slightly, in the whole or at certain localities, slow evolution of dissolved air will follow which may be undetectable by ordinary observation. If the pressure drops suddenly, air evolution is accelerated and bubble growth may become observable before the rising bubbles reach the top surface of the water. Such bubbles will appear first near solid surfaces, not only because the hydraulic pressure usually is the lowest at such spots, but also because the chemical bonds there are weaker than the cohesion between the water molecules. If the pressure drops below the vapor pressure of the liquid, evolution of vapor into the nuclei will begin at a relatively high rate, thus resulting in fast bubble growth. Frequently this is regarded as the beginning of incipient cavitation.

With liquids other than water, or quiescent water, or with pure denucleated water flowing in a chemically clean container, the walls of which are well wetted by the water (hydrophilic material), cavitation will not, as a rule, begin at vapor pressure. A much lower pressure is required to start microcavitation and a higher than vapor pressure is sufficient to turn microcavitation into macrocavitation, both depending on exterior circumstances.

This investigation was done in 1948 and 1949 at the sponsorship of the Air Materiel Command U. S. Air Force.

Electrical Conductivity Method for Measuring Self-Diffusion of Metals

J. H. DEDRICK*
Sylvania Electric Products, Inc., Bayside, New York

AND

G. C. KUCZYNSKI
Bogota, Colombia

(Received May 29, 1950)

A method has been proposed to measure self-diffusion in metals by measurement of the contact conductance of an interface between two metallic spheres or hemispheres. By combining Kuczynski's equation for self-diffusion with Bowden and Tabor's equation for the contact conductance between two metallic spheres a relationship has been obtained between contact conductance and the time of heating of the spheres. Thus by measurement of junction conductance as a function of time of heating, for different temperatures, diffusion coefficients and the heat of self-diffusion can be evaluated in the normal manner. A similar relationship has been obtained for two spheres or hemispheres initially pressed together to give a known contact conductance before heating.

ONE of the authors[1] developed a theory of sintering, verified by extensive experimental work, showing that the initial bonding of like metallic particles upon heating is a volume diffusion process. The relationship developed was

$$X^6/a^2 = 40\sigma\delta^3/kT \cdot Dt, \qquad (1)$$

where X is the radius of the interface between the particles (assumed to be circular) a the original radius of the particles, σ the surface tension of the material, δ the interatomic jump distance, k the Boltzman constant, T the absolute temperature, t the time, and D the coefficient of self-diffusion. This relationship shows that the fifth power of the interface radius between the particles is proportional to the time of sintering. Values of D, D_0 (frequency factor) and Q (heat of self-diffusion) obtained in this manner on copper and silver agree well with those obtained by radioactive tracer methods.

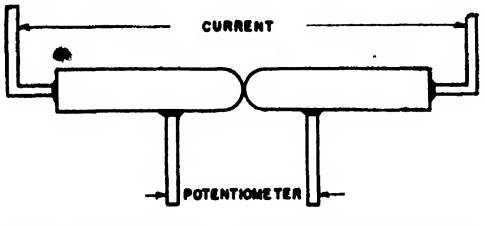


FIG. 1. Method of measuring junction conductance.

The technique used to measure the interface radius as a function of time for a given temperature was to sinter spherical metallic particles to a base plate and to measure the radius by metallographic means. This method was later changed by using wire wound around a mandrel of the same material. The use of blown-up low magnification photographs has also been resorted to for measuring contact radii of particles.[2] In this case a single close packed layer of particles was used.

This article relates to another method which might be used to measure heats of self-diffusion and self-diffusion coefficients based upon a relationship between Eq. (1) and the contact conductance of an interface between two metallic spheres or hemispheres.

* Now at Reynolds Metals Company, Louisville, Kentucky.

[1] G. C. Kuczynski, "Self-Diffusion in sintering of metallic particles," Metals Trans. 169–178 (February, 1949).

[2] J. H. Dedrick and A. Gerds, J. App. Phys. 20, 1042–1044 (1949).

Bowden and Tabor[3] in their work on measurement of the area of contact between stationary surfaces, have shown that the junction conductance (C_j) between two metallic spheres of radius a, or crossed cylinders, joined at the contact by a circle of radius X, where X is small compared to a, is given by

$$C_j = 2\lambda X, \tag{2}$$

where λ is the specific conductance of the bulk material. Using crossed cylinders of silver (0.90-cm radius) and increasing the contact area by varying the load, they found a direct proportionality between conductance and contact radius up to a value of the latter of 0.08 cm which is approximately 18 percent of the cylinder radii.

By combining (1) and (2) the following relationship is obtained between junction conductance and time of heating of two spheres or two hemispheres in contact with each other:

$$\frac{(C_j/2\lambda)^5}{a^2} = 40\sigma\delta^3/kT \cdot Dt. \tag{3}$$

Thus from measurements of junction conductance as a function of time of heating, for different heating temperatures, D values could be obtained and D_0 and Q evaluated in the normal manner. This method makes it unnecessary to resort to metallographic work, and one set of specimens could be used for different times of sintering at a given temperature. The possibility of continuous recording of junction conductance as a function of time of heating also presents itself. A schematic sketch for the measurement of the junction conductance between two hemispherically ended specimens is shown in Fig. 1.

The discussion up to this point has been based upon the assumption that the initial point junction conductance will be very low and thus neglected. However, if this is not so, a different mode of attack would have to be resorted to and it appears that the simplest approach would be to initially press the spheres or hemispheres together with a certain force and thus produce an initial junction radius X_0, as shown in Fig. 2, from which the initial junction conductance C_j^0, prior to heating, could be measured.

In this case, it would be necessary to use the following

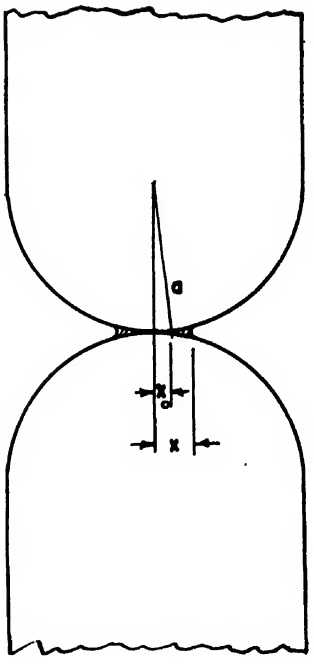


FIG. 2. Sketch of contact junction between two pressed hemispherical ended rods.

relationship based upon the geometry of the figure:

$$\frac{X^5 - X_0^5}{5} - 2X_0^2\frac{X^3 - X_0^3}{3} + X_0^4(X - X_0) = At, \tag{4}$$

where $A = (40\sigma\delta^3)/kT(a - X_0)^2D$, and which reduces to (1) when X_0 equals zero. If X is not very much larger than X_0 (4) can be written as $(4/3)X_0^2(X - X_0)^3 = At$, or

$$X = X_0\left\{1 + \left(\frac{3A}{4X_0^5}\right)^{\frac{1}{3}} t^{\frac{1}{3}}\right\}, \tag{5}$$

and by combining (2) and (5), one obtains

$$C_j = C_j^0\left\{1 + \left(\frac{24A^5}{(C_j^0)^5}\right)^{\frac{1}{3}} t^{\frac{1}{3}}\right\},$$

or

$$C_j = C_j^0\{1 + Kt^{\frac{1}{3}}\}, \tag{6}$$

where K is a constant for a given material, for a given temperature of heating and a known initial junction conductance. Thus as in the case of initially unpressed spheres or hemispheres, measurements of junction conductance as a function of time, for different heating temperatures will allow for determination of D values and evaluation of D_0 and Q.

[3] F. P. Bowden and D. Tabor, "The area of contact between stationary and moving surfaces," Proc. Roy. Soc. London 169A, 391 (1939).

A Simple Interferometric Test for Conical Flow

J. H. Giese, F. D. Bennett, and V. E. Bergdolt
Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland
(Received May 3, 1950)

In conical flows the velocity, pressure, and density do not vary on lines through the vertex of the cone. In this paper it is shown that for interferograms of general conical flows $\delta(y, z)/z=f(y/z)$, where y and z are any set of Cartesian coordinates with origin at the image of the vertex of the cone, and δ is the fringe shift at y, z. Thus, for strictly conical flow, a graph of $\delta(y, z)/z$ *versus* y/z should be a single curve. This suggests a test for approximate conicity that requires very little computation. This test is applied to interferograms obtained from a number of approximately axisymmetric flows at various Mach numbers about cone-cylinders in free flight. Plotted fringe shift data from the region near the nose fall into a narrow band, an indication of approximate conicity. They also closely check the corresponding theoretical fringe shift curve calculated for Taylor-Maccoll flow.

I. INTRODUCTION

INTERFEROMETRY offers an extremely attractive technique for investigation of gas flows. When it can be employed, an extensive region of the flow field can be surveyed almost instantaneously without disturbing the flow. Accordingly, it is well suited for examining supersonic flows about objects in free flight, as well as those about stationary models in wind tunnels. If the flow is plane or axisymmetric, the density distribution can be computed from fringe shifts observed on an interferogram, though the latter case involves formidable amounts of calculation. At present the quantitative interpretation of interferograms in less symmetrical flows, such as those about yawing bodies of revolution or about more general shapes, seems to be limited to the comparison of measured and theoretical values of fringe shifts. The calculation of theoretical fringe shifts for general flows with little or no symmetry is difficult and tedious at best, and the results depend so strongly on the shape of the body and its orientations relative to the incident stream and the interferometer beam that it seems impossible to reach interesting general conclusions without actually performing the computations. In the important special case of conical flow, however, one can easily make a useful deduction about the functional form of the theoretical fringe shift. Before this is derived, a brief digression on conical flow seems apropos.

A steady flow of a fluid with no viscosity or heat conduction is said to be *conical* if the velocity components, pressure, and density do not vary along lines emanating from some fixed point, which will be called the *vertex*.[1–3] The use of these terms is justified intuitively by applying similarity considerations to flows about one infinite nappe of a general cone placed in a uniform supersonic incident stream. If the vertex of the flow is placed at infinity, plane and swept-back plane flows result.[1] Owing to the remarkable geometrical structure of conical flow, its theory is considerably simpler than that of more general three-dimensional flows and has been developed to a much greater extent. The best known examples of non-linearized conical flows with vertices at finite locations are Taylor-Maccoll flow about a non-yawing right circular cone;[2–4] Prandtl-Meyer expansion around an edge;[3] and the flow over a wedge with attached plane shock wave.[3] The linearized theory has been applied by numerous authors to determine flows over objects ranging from elliptic cones to delta-wings.

The great theoretical interest and value of conical flows make it essential to determine how closely they correspond to flows that can actually be produced in real fluids. The commonest type of test is a measurement of the pressure distribution on the surface of a conical model in a supersonic wind tunnel.[4] Under good test conditions, the pressure is found to be reasonably constant on lines through the vertex of the model, and the observed pressures check the available linearized or non-linearized theoretical values to within a few percent. For flows about cones of revolution and wedges, measured and theoretical shock wave angles have also been compared.[4] As far as interferometry is concerned, Ladenburg[5] has found contours of constant density in the flow about a cone of revolution in an approximately axisymmetric jet. At large enough distances from the vertex of the cone, his measured densities are very nearly constant on radial lines through the vertex and check the theoretical Taylor-Maccoll values to well within two percent.

In this paper a test for conicity, based directly on interferogram fringe shifts, will be developed and applied to data obtained from conical regions in flows about bodies in free flight.

II. THEORETICAL CONSIDERATIONS

A. Functional Form of Fringe Shift

The following reasoning leads to a simple method for interpreting interferograms of general conical flows. Let

[1] A. Busemann, Luftfahrtforschung 12, 210–220 (1935).
[2] A. Busemann, Zeits. f. angew. Math. Mech. 9, 496–498 (1929).
[3] A. Ferri, *Elements of Aerodynamics of Supersonic Flows* (New York, 1949), Chapters 2, 3, 12.
[4] G. I. Taylor and J. W. Maccoll, Proc. Roy. Soc. A139, 278–311 (1933).
[5] Ladenburg, Winckler, and Van Voorhis, Phys. Rev. 73, 1359–1377 (1948).

the vertex of the flow be at the origin of a system of rectangular coordinates. Assume that the interferometer beam that traverses the flow is parallel to the x-axis, while the photographic plate is normal thereto, so the projections of the y- and z-axes onto the plate define rectangular coordinates with origin at the image of the vertex. At (y, z) on the plate, in the absence of magnification, the fringe shift $\delta(y, z)$ is defined by

$$\delta(y, z) = (K/\lambda) \int_{a(y,z)}^{b(y,z)} [\rho(x, y, z) - \rho_u] dx, \qquad (1)$$

where ρ_u is the density of the undisturbed fluid, assumed constant; $\rho(x, y, z)$ is the actual density at (x, y, z); K is the Gladstone-Dale constant for the fluid, i.e., the ratio of small increments of index of refraction to the corresponding small increments of density; λ the vacuum wave-length of the light used to form fringes; and the limits of integration $a(y, z)$ and $b(y, z)$ are the x coordinates of the points where the light ray through (y, z) crosses the conical shock wave, outside of which $\rho(x, y, z) = \rho_u$. For any $t > 0$

$$\delta(ty, tz) = (K/\lambda) \int_{a(ty,tz)}^{b(ty,tz)} [\rho(x, ty, tz) - \rho_u] dx.$$

Since the flow is conical, $\rho(x, ty, tz) = \rho(x/t, y, z)$, and since (ty, tz), (y, z), and $(0, 0)$ are collinear, $a(ty, tz) = ta(y, z)$, and $b(ty, tz) = tb(y, z)$. Hence

$$\delta(ty, tz) = (K/\lambda) \int_{ta(y,z)}^{tb(y,z)} [\rho(x/t, y, z) - \rho_u] dx.$$

Let $\xi = x/t$ to obtain

$$\delta(ty, tz) = t(K/\lambda) \times \int_{a(y,z)}^{b(y,z)} [\rho(\xi, y, z) - \rho_u] d\xi = t\delta(y, z), \qquad (2)$$

i.e., $\delta(y, z)$ is homogeneous of degree one. Now let $t = 1/z$, to find $\delta(y, z)/z = \delta(y/z, 1) = f(y/z)$ for some function f. Hence, *for conical flow a plot of* $\delta(y, z)/z$ *versus* y/z *should be a curve*,* a conclusion that can also be obtained from similarity considerations. To test a flow qualitatively for conicity by means of an interferogram, it suffices to measure $\delta(y, z)$ as a function of y on a series of traces $z =$ constant, and to observe how closely the points y/z, $\delta(y, z)/z$ come to falling on a curve. Since this procedure involves very little computation with the measured data, it can be used advantageously in either the axisymmetrical or non-axisymmetrical conical case.

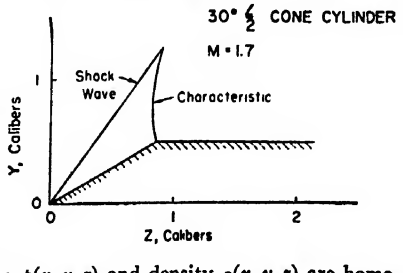


FIG. 1. Theoretical conical flow region.

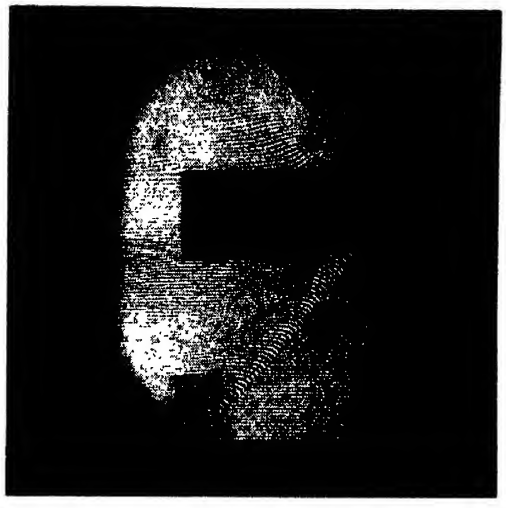


FIG. 2. Interferogram of cal. 30, 30° half-angle cone cylinder, $M = 1.7$, $p = 1$ atmos., room temperature.

B. Flows Containing Conical Regions

Let us turn, now, to the problem of producing an approximately conical flow to which to apply this test. Earlier remarks suggest the use of a conical model. If the flow is supersonic, the effects of the disturbance produced by ending the cone at some finite distance from its vertex will merely be propagated downstream without influencing conditions near the vertex. Thus at supersonic speeds high enough to assure the presence of an attached conical shock wave, the flow about a conical model in a wind tunnel or about a projectile with a conical nose should contain a finite region of conical flow bounded by the following surfaces as indicated in Fig. 1: a conical segment of the shock wave; the conical part of the model's surface; and a characteristic surface through the curve at which the physical cone ends. In particular, if the model is axisymmetric and does not yaw, then near its vertex there will be a region of Taylor-Maccoll flow.

C. Taylor-Maccoll Fringe Shifts

In order to provide a quantitative comparison for the measurements of flows about cone-cylinders that will be presented later, theoretical fringe shift curves will be required. These can be computed as follows: In conical flow that is axisymmetric about the z-axis,

$$\rho = \rho(\theta), \quad \tan^2\theta = (x^2 + y^2)/z^2,$$

$$\lambda\delta(y, z)/2K\rho_u z = \int_0^b (\rho(\theta)/\rho_u - 1) dx/z$$

* Since the pressure $p(x, y, z)$ and density $\rho(x, y, z)$ are homogeneous of degree zero, plots of $(p - p_u)/\frac{1}{2}\rho_u q_u^2$ or of ρ/ρ_u (where ρ_u and q_u are the density and speed of the undisturbed flow) *versus* y/z on a surface $x/z = f(y/z)$ should also be curves. Of course, such plotting is equivalent to testing for constancy of p and ρ on lines through the vertex.

where $b^2=z^2\tan^2\theta_w-y^2$; $\tan\theta_s \leqq y/z \leqq \tan\theta_w$; and θ_s and θ_w are the semi-vertex angles of the cone and shock wave, respectively. Integrate the constant term in the integrand, and transform the variable from z to θ to obtain

$$\lambda\delta(y,z)/2K\rho_u z=\int_{\tan^{-1}y/z}^{\theta_w}[\rho(\theta)/\rho_u]\tan\theta\sec^2\theta \times(\tan^2\theta-y^2/z^2)^{-\frac{1}{2}}d\theta-(\tan^2\theta_w-y^2/z^2)^{\frac{1}{2}}. \quad (3)$$

Behind the shock wave, $p/p_0=(\rho/\rho_0)^\gamma$, where p is the local static pressure; p_0 and ρ_0 are the stagnation pressure and density, respectively; and γ is the ratio of the specific heat at constant pressure to that at constant volume. The local speed of sound, a, is defined by $a^2=dp/d\rho=\gamma p/\rho$. Then $\rho/\rho_0=(a/a_0)^{2/(\gamma-1)}$, where a_0 is the stagnation speed of sound. Similarly, $\rho_u/\rho_{01}=(a_u/a_{01})^{2/(\gamma-1)}$, where ρ_{01} and $a_{01}=a_0$ are the stagnation density and speed of sound, respectively, ahead of the shock wave. By Bernoulli's equation, $\frac{1}{2}q_u^2+a_u^2/(\gamma-1)=a_0^2/(\gamma-1)$, where q_u and a_u are the undisturbed speeds of flow and of sound, respectively. Then $(a_0/a_u)^2=1+\frac{1}{2}(\gamma-1)M^2$, where $M=q_u/a_u$ is the Mach number of the undisturbed flow. Thus

$$\rho/\rho_u=(\rho_0/\rho_{01})[1+\tfrac{1}{2}(\gamma-1)M^2]^{1/(\gamma-1)}(a(\theta)/a_0)^{2/(\gamma-1)}.$$

Since θ_w, ρ_0/ρ_{01}, and $a^2(\theta)$ (in velocity units such that $a_0=1$) have been tabulated as functions of θ_s, M, and θ by Kopal[6] by numerical integration $\lambda\delta(y,z)/2K\rho_u z$ can be determined as a function of y/z. The singularity of the integrand at $\theta=\tan^{-1}y/z$ can easily be avoided by splitting the integration into two parts: one with respect to θ from some convenient $\theta_0>\tan^{-1}y/z$ to θ_w; the other with respect to x from 0 to $(z^2\tan^2\theta_0-y^2)^{\frac{1}{2}}$. In actual use, the values of $\delta(y,z)/z$ so computed will have to be multiplied by a magnification factor appropriate to the experimental optical system. It must also be mentioned, as an aid to plotting, that one can easily show that $d[\delta(y,z)/z]/d(y/z)$ is infinite at $y/z=\tan\theta_w$.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Flows to be tested for conicity were produced by firing cal. 0.30 cone-cylinders at an absolute pressure of one atmosphere in the range described Bergdolt *et al.*[7] During these firings interferograms of which Fig. 2 is typical were obtained from a Mach-Zehnder interferometer with two-inch plates, used in conjunction with a rotating mirror which approximately stopped the images of the bullets. The light employed was from the 5460A line of mercury, produced during a flash of about 2 microseconds duration from a BH6 mercury tube. Other relevant data are appended to the plots of smoothed fringe shift measurements shown in Figs. 3 to 6.

In general, the points plotted in Figs. 3 to 6 lie in

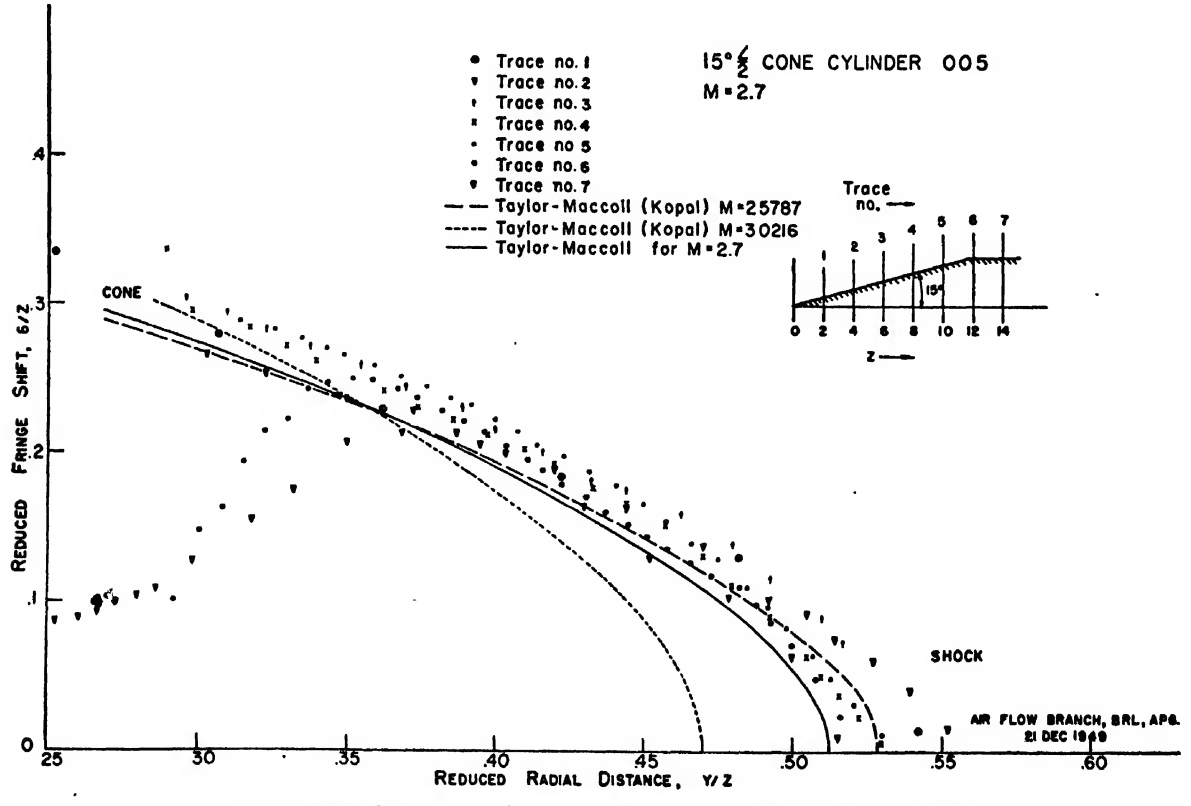


FIG. 3. Conical flow test plot showing departure from conicity in expansion region.

[6] Z. Kopal, *Tables of Supersonic Flow around Cones* (Cambridge, Mass., 1947).
[7] Bergdolt, Greenwald, and Sleator, Phys. Rev. 76, 879 (1949).

rather narrow bands, which encourages the belief that these flows came fairly close to being conical. In Fig. 3, for traces beyond the shoulder of the cone-cylinder, the points near the shock wave lie close to the conical flow band, while those near the body fall far outside. This verifies that the region of conical flow is really qualitatively of the theoretical form shown in Fig. 1. Actually this is also clear by direct inspection of Fig. 2, where the trace of the characteristic surface that bounds the region of conical flow is plainly discernible. Since the vertex of a conical flow is a singularity of the theoretical flow field, close to which the velocity and other properties of the flow have large gradients, even for a perfectly pointed model one would expect the behavior of a real viscous fluid to differ considerably there from that of the theoretical perfect fluid. The actual body, however, must have a slightly blunted nose, so the shock wave must be detached and doubly curved rather than conical. *A fortiori*, one would expect deviations from conicity near the tip of the body; however, the present measurements do not seem to exhibit such effects. In Figs. 3 and 6, points of the first trace bear no particu larly significant relation to those of others farther downstream. In Figs. 4 and 5, on the other hand, points of the first traces are systematically displaced above the rest. The magnitude of this displacement is consistent with the assumption that an error of the order of 0.2 mm was made in determining the location of the vertex on the interferogram. Such errors can arise, for example, if the vertex falls on a dark fringe. For an interferogram with well defined vertex, no such displacements are observed, and all of the plotted points fall within a narrow band, as in Fig. 6.

Estimates of the erratic experimental error $\Delta\delta$ in fringe shift indicate that $\Delta\delta \leqslant 0.05$ for all of the measurements reported here. For a given trace, percentage error in δ will vary from a large value near the shock wave to a minimum at the projectile body. Fringe shift generally increases from the tip of the projectile toward the rear; so percentage erratic error is higher for low trace number than for high.

For the data represented in Fig. 6 $\Delta\delta \leqslant 0.05$ indicates error for trace No. 1 of 18 percent or less at the projectile boundary. This upper bound falls to 10 percent for trace No. 2, 6 percent for trace No. 3, 4 percent for trace No. 5, and less than 3 percent for the remaining traces. These estimates apply to the δ/z values as well; for the erratic error in location of z is of the order of a few tenths percent even for smallest z.

IV. COMPARISON WITH TAYLOR-MACCOLL FLOW

A quantitative check of observed fringe shifts requires a corresponding theoretical flow model. In the present case the obvious choice, Taylor-Maccoll flow, must be justified by inquiring how well the fundamental assumptions of uniformity ahead of the shock wave, inviscidness, steadiness, and axisymmetry are satisfied. As far as viscosity is concerned, reasonably high

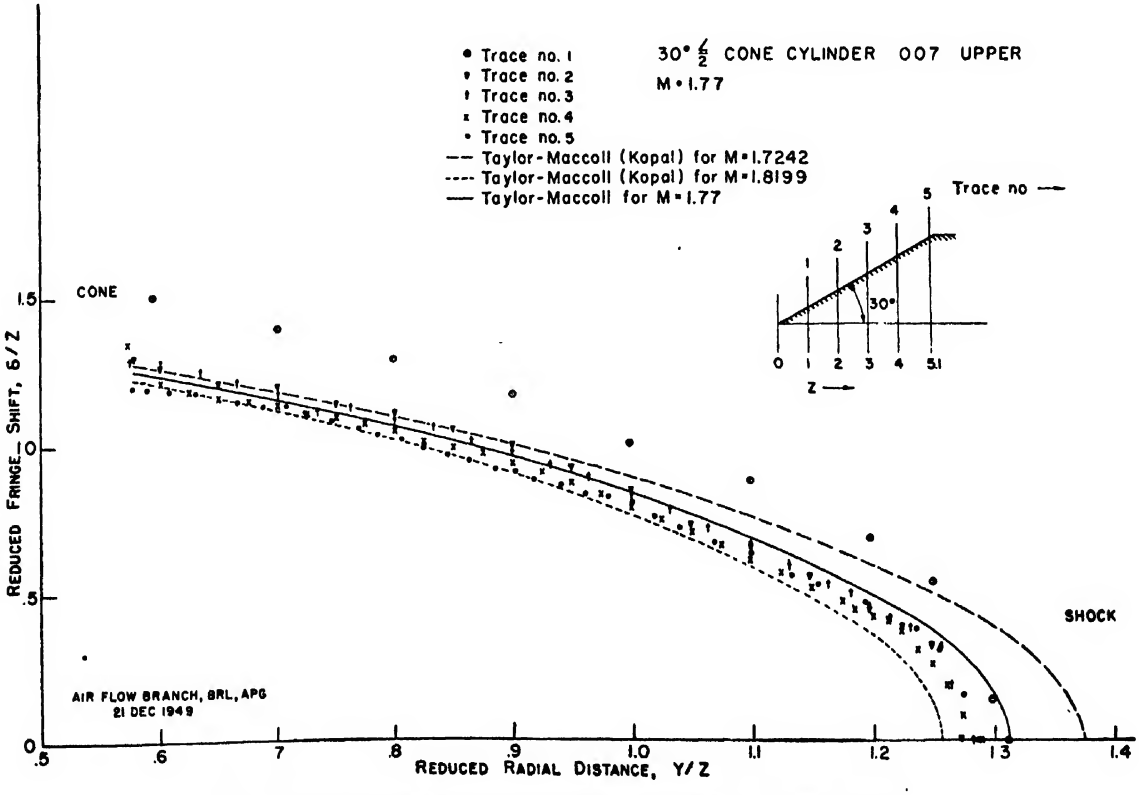


FIG. 4. Conical flow test plot of upper half of interferogram shown in Fig. 2.

Reynolds numbers, on the order of 3×10^5 based on the altitude of the cone, lead one to expect only very thin boundary layers. Actually, no boundary layer effects are readily perceptible on the interferograms, since the appearance of the fringes near the cone discloses nothing comparable to the hooks in Bershader's fringes in a well developed boundary layer.[8] Though the body is spin stabilized, on the cone the component of velocity due to spin is at most 4 percent of the forward velocity and thus will certainly not contribute significantly to viscosity effects. Unsteadiness may be produced by the deceleration and oscillation of the bodies. A rough estimate of the deceleration is 700 ft./sec.^{2}. At a velocity of 2000 ft./sec. the relative change in velocity is approximately $\Delta v/v=0.7\times10^{-5}$ during the time required for passage through the conical region, which is less than 20 microseconds outside the boundary layer. A crude upper bound on the frequency of oscillations in the flow can be derived from the fact that the fringes are steady and sharply defined during a 2 microsecond exposure. As for oscillations of the body, the precessional and nutational components of its motion have frequencies of less than 20 and 200 cycles/sec., respectively. The corresponding periods are quite long by contrast with the time of passage through the conical region. The fact that the body is in free supersonic flight in still air assures one that the flow ahead of the shock will be uniform; this is confirmed by the straightness of the undisturbed fringes. Finally, as for axisymmetry, though small yaws are inevitable, for conventional projectiles at normal pressures it is difficult to produce large yaws. Actually, measurement of the interferograms shows that the inclination of the body's axis to the trajectory is about half a degree in this projection. As a check on the influences of such small yaws and of the differences in fringe geometry, results from the upper and lower halves of the same interferogram are shown in Figs. 4 and 5.

Granting that conditions favorable for approximately Taylor-Maccoll flow occur in the present case, there will still be errors in the data. Two types will be considered here. First, in the calculation of Taylor-Maccoll fringe shifts by the method of Section II, it is assumed that the interferometer beam is normal to the axis of symmetry. Now the dynamic history along the trajectory cannot easily be predicted with great precision; moreover, it will vary from one round to the next with the geometrical and physical properties of the body. If for this reason the curvature of the trajectory and the oscillations of the body are neglected, the best approximation to the axis of symmetry is that of the gun barrel. The inclination of the interferometer beam relative to this axis was probably within two degrees of orthogonality. Calculations show that the effects of such an error in inclination are negligible. Secondly, no instrumentation

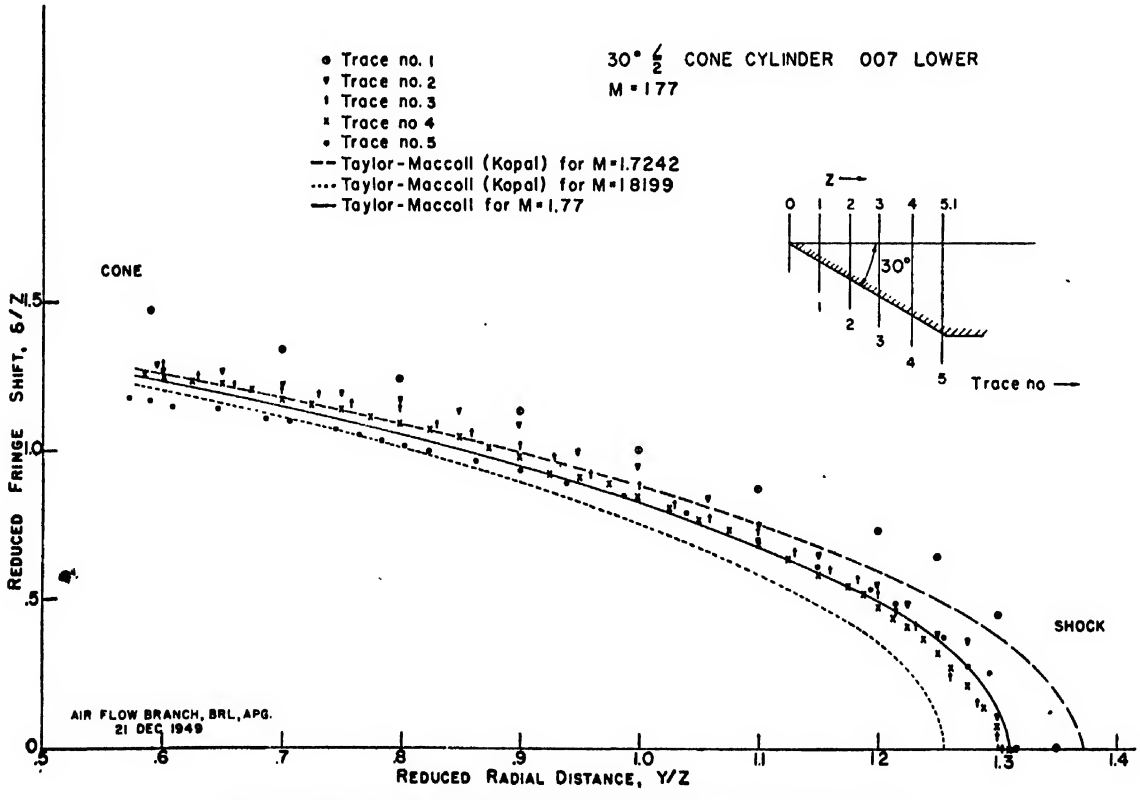


FIG. 5. Conical flow test plot of lower half of interferogram shown in Fig. 2.

[8] D. Bershader, Rev. Sci. Inst. 20, 260–275 (1949).

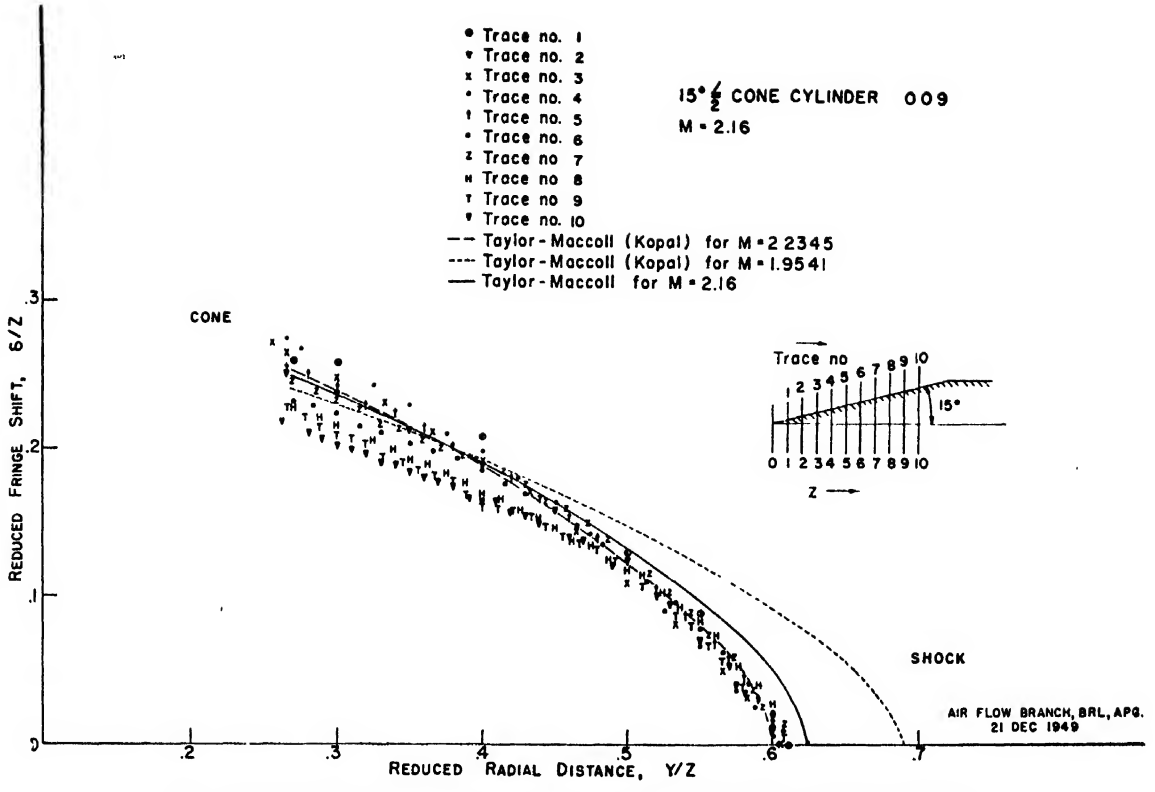


FIG. 6. Conical flow test plot for cal. .30, 15° half-angle cone cylinder—best data available.

for directly measuring velocities was available for the present firings. As a first approximation, the velocity was taken to be the average observed in firings of the same type of body with the same weight of powder in other ranges equipped with counter chronographs. A more refined estimate of the Mach number was based on the theoretical shock wave angle in Taylor-Maccoll flow, which is justified, *a posteriori*, by the agreement that will be found between the experimental data and the theoretical curves.

Figures 4 and 5 show fringe shift curves for $M=1.72$ and 1.82, calculated with $\gamma=1.4$ for Taylor-Maccoll flows tabulated in reference 6. The curve for $M=1.77$ was obtained by linear interpolation between the other two. In general, the agreement between the experimental data and theoretical curves is good, the poorer agreement in Fig. 5 being caused, possibly, by errors in the location of the axis and vertex. The scatter of the points at the shock ends of the curves may be partly due to curvature of the shock wave and partly to measurement errors, the fringe shifts being smallest there. At the cone end of the curves, the spread could conceivably originate in the boundary layer, since the traces farthest downstream, in which coordinate errors are least, seem in some instances to be displaced toward lower values.

From the quantitative estimates of error given in the last paragraph it appears that the scatter of points near the projectile body for traces posterior to No. 5 should be less than 3 percent from a mean curve. Inspection of the plotted points in Fig. 6 confirms this view. It is noteworthy that as trace number increases these latter traces fall increasingly below the Taylor-Maccoll curve for $M=2.16$ as well as below the mean curve for the first 5 traces. Translation of the data for any trace to bring the shockwave point into coincidence with that of the calculated Taylor-Maccoll curve does not improve matters much, since near the cone the latter traces still fall below the theoretical curve by several percent. We see then that the data suggests two types of deviation from conical flow (1) deviation from the Taylor-Maccoll theoretical flow of the latter traces and (2) deviation of the latter traces from the conical flow defined by the mean curve of the former traces. It is possible that these deviations can arise in the boundary layer as suggested above; however further study is required to obtain more than tentative conclusions to this effect.

Theory of Collinear Antennas

RONOLD KING
Cruft Laboratory, Harvard University, Cambridge, Massachusetts
(Received May 1, 1950)

If N collinear antennas are individually center driven, the only transmission line in the neutral plane is the one for a central unit if N is odd. Since lines that are not in a neutral plane constitute antennas, such an array is not practically useful.

For a single driven unit with symmetrical, parasitic, collinear elements, the driving transmission line is in the neutral plane and may be replaced by an equivalent generator and terminal-zone network. An array with two parasites is analyzed. Since the currents in the halves of the three units are not alike, three simultaneous integral equations occur. By making reasonable assumptions these are reduced to the single equation of the isolated antenna but with a different kernel. Approximate expressions are obtained for the currents and for the mutual and self-impedances of the elements.

The collinear antenna with outer elements driven from the central one by phase-reversing stubs is studied. For three elements components of the currents in the several possible modes are obtained with different degrees of approximation. The in-phase currents in the antennas are greatest, and are obtained with considerable accuracy. The currents in the outer units are not symmetrical, but with elements of length $\lambda/2$ the odd components of the currents are shown to be negligible.

1. INTRODUCTION AND GENERAL FORMULATION OF THE PROBLEM

THE analysis of the two individually driven collinear antennas in Fig. 1a is complicated by the fact that the axes of the transmission lines are not in neutral planes of the electromagnetic field. As a result, the forces acting on currents and charges in the two conductors of each line are not equal and opposite even at distances from the antennas that are great compared with the normal terminal zone[1] near each junction. Therefore, the transmission lines are not balanced, and the codirectional currents on each pair of conductors (or on the outside of the shield if a coaxial line is used) contribute significantly to the electromagnetic field at distant points. The actual radiating system in Fig. 1a is not limited to the two collinear antennas but includes the transmission lines. Clearly, such a system is not practically useful since its circuit and field properties depend on the length and orientation of the feeder lines.[2]

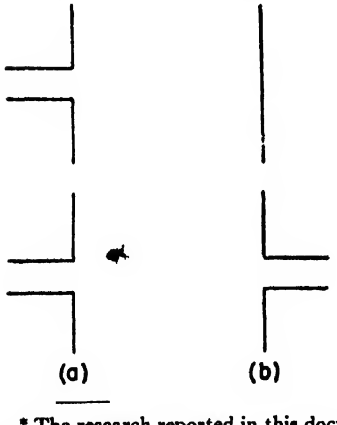


FIG. 1. Arrays of two collinear antennas.

The same difficulty exists if one of the two elements is parasitic as in Fig. 1b; or if three individually center-driven elements are used as in Fig. 2a. On the other hand, the array in Fig. 2b, in which only the central one of an odd number of collinear elements is directly driven, does not have the same serious disadvantage. The single transmission line is in the neutral plane of the array, so that unbalanced currents are not excited and radiation from the line is negligible.

The complication introduced by the two-wire line in Fig. 2b differs in no essential way from that already resolved for a single, center-driven antenna.[1] The presence of the two parasitic elements has no effect on the coupling between antenna and line since this is confined to short distances from the line-antenna junction of the order of magnitude of ten times the line spacing. It follows that the same terminal-zone network designed for a single isolated center-driven antenna may be used in conjunction with the antenna, treated as though driven by a discontinuity in scalar potential. This and following sections are concerned with the analysis of a three-element collinear array arranged and numbered as in Fig. 2b but with the central unit driven by a discontinuity in scalar potential, i.e. by a two-wire line with negligible line spacing. The analysis is carried out conveniently in the following parts:

(1) *The Symmetrical Problem:* The elements of the array are individually center-driven by discontinuities in scalar potential given by

$$V_{10}=V_{1}{}^{s}=V^{s};\quad V_{20}=V_{30}=V_{2}{}^{s}=f^{s}V^{s}. \qquad (1.1a)$$

$V_{2}{}^{s}$ is so chosen that the currents at the centers of the identical outer elements are equal to the current at

* The research reported in this document was supported in part by the Navy Department (ONR), the Signal Corps of the U. S. Army, and the U. S. Air Force, under contract N5-ori-76, T.O. 1/.

[1] R. King, J. App. Phys. 20, 832–850 (1949).

[2] An approximate analysis of the two-element collinear array in which no account is taken of the practical problem of driving the array is that of C. W. Harrison, Jr., Proc. I.R.E., 33, 398–408 (1945). Zeroth-order mutual impedances given by Harrison coincide with those derived in this paper. The general solutions differ in that Harrison did not include the effect of the coupled antenna in the kernel of the integral equation, nor did he take account of the distribution of current in deriving his expansion parameter.

the center of the middle element. That is, so that

$$I_2^s(z_2=0)=I_3^s(z_3=0)=I_1^s(z_1=0) \qquad (1.1b)$$

where z_1, z_2, and z_3 are measured from the centers of the elements indicated by the subscripts.

(2) *The Antisymmetrical Problem:* The three elements of the array are individually center-driven by

$$V_{10}=V_1^a=V^a; \quad V_{20}=V_{30}=-V_2^a=-f^aV^a \qquad (1.2a)$$

V_2^a is chosen that the currents at the centers of the outer elements are equal in magnitude but opposite in direction to the current at the center of the middle unit. That is, so that

$$I_2^a(z_2=0)=I_3^a(z_3=0)=-I_1^a(z_1=0). \qquad (1.2b)$$

(3) *The Parasitic Array:* By superimposing the solutions for the currents in problems (1) and (2) in the special case when

$$V_2^a=V_2^s \quad \text{or} \quad f^aV^a=f^sV^s, \qquad (1.3a)$$

the effective driving voltages and driving-point currents are

$$\begin{cases} V_{10}=V^s+V^a=V^a(1+f^a/f^s)=V^a(1+k), & (1.3b) \\ I_1(z_1)=I_1^s(z_1)+I_1^a(z_1), & (1.3c) \end{cases}$$

and

$$V_{20}=V_2^s-V_2^a=0, \qquad (1.3d)$$

$$I_2(z_2)=I_2^s(z_2)-I_2^a(z_2)$$

$$=I_{\text{even}}^s(z_2)+I_{\text{odd}}^s(z_2)-I_{\text{even}}^a(z_2)-I_{\text{odd}}^a(z_2). \qquad (1.3e)$$

Since $V_{20}=0$, the outer units are parasitic. Moreover, since the currents in the outer units are not symmetrical with respect to the centers of the respective elements,

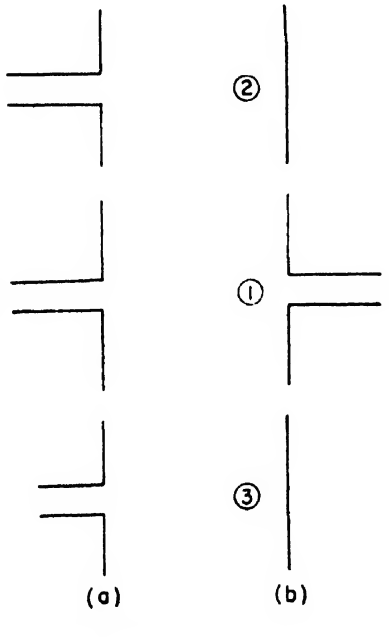


FIG. 2. Arrays of three collinear antennas.

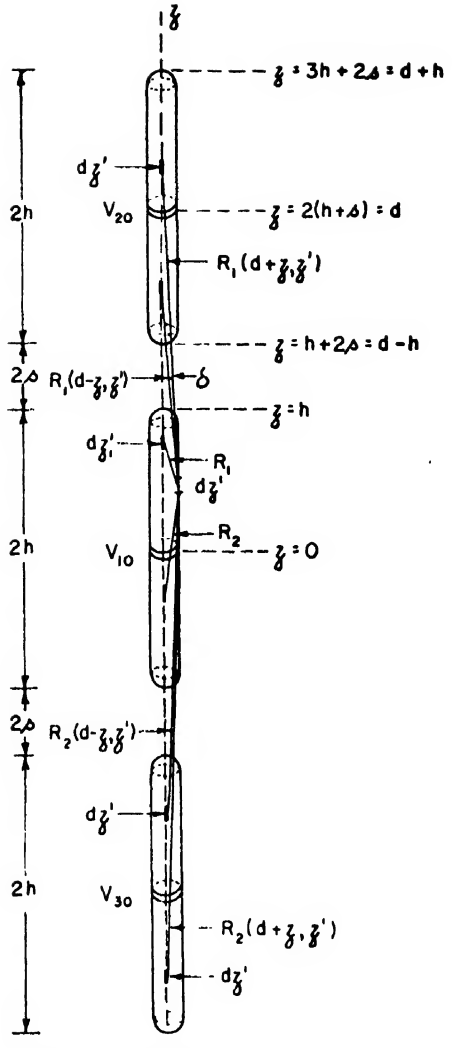


FIG. 3. Collinear array with slice generators.

the symmetrical and antisymmetrical currents are each the sum of even and odd components that must be determined separately.

(4) *The General Problem:* If the following condition is satisfied in place of (3a) or (3d):

$$V_{20}=V_2^s-V_2^a=-I_2(z_2=0)Z_L, \qquad (1.4)$$

the solution with each outer unit center-loaded with a lumped impedance Z_L is obtained. Note that $I_3(-z_3)=I_2(z_2)$.

The particular array studied in this and following sections is in Fig. 3. The three antennas are identical, each of half-length h and radius a. The conditions

$$a\ll h; \quad 2\pi a/\lambda_0=\beta_0 a\ll 1 \qquad (1.5)$$

are postulated. The distance between centers of the antennas is d, the distance between adjacent ends is $2s$ so that $d=2(h+s)$. The origin of a cylindrical system of coordinates r, θ, z, is at the center; the z-axis coin-

cides with the axes of the cylinders. The centers of the three units are at $z=0, \pm d$. Each unit is center-driven by discontinuities in scalar potential. Referring to Fig. 3 the conditions for the symmetrical drive are as in (1.1a), i.e.,

$$V_{10}=\lim_{\delta\to 0}[\phi(\delta)-\phi(-\delta)]=V_1^s=V^s \tag{1.6a}$$

$$V_{20}=\lim_{\delta\to 0}[\phi(d+\delta)-\phi(d-\delta)]=V_2^s=f^sV^s \tag{1.6b}$$

$$V_{30}=\lim_{\delta\to 0}[\phi(-d+\delta)-\phi(-d-\delta)]=V_2^s=f^sV^s \tag{1.6c}$$

where V_2^s or f^s is so chosen that (1.1b) is satisfied, i.e.,

$$I_3^s(-d)=I_2^s(d)=I_1^s(0). \tag{1.6d}$$

Similarly, the conditions for the antisymmetrical drive are as in (1.2a) with V_2^a or f^a chosen so that (1.2b) is satisfied, i.e.,

$$I_3(-d)=I_2(d)=-I_1(0). \tag{1.7}$$

Since the array is symmetrical with respect to its center, for both sets of driving conditions,

$$I(-z)=I(z); \quad q(-z)=-q(z), \tag{1.8a}$$

$$A(-z)=A(z); \quad \phi(-z)=-\phi(z). \tag{1.8b}$$

The subscript z is omitted on current and vector potential, since only z-components are involved. The vector and scalar potentials, evaluated at $r=a$ on the surfaces of the conductors, are

$$A(z)=\frac{1}{4\pi\nu_0}\left(\int_{-d-h}^{-d+h}+\int_{-h}^{h}+\int_{d-h}^{d+h}\right)I(z')K_1(z,z')dz'$$

$$=\frac{1}{4\pi\nu_0}\left(\int_0^h+\int_{d-h}^{d+h}\right)I(z')K(z,z')dz' \tag{1.9a}$$

$$\phi(z)=\frac{1}{4\pi\epsilon_0}\left(\int_0^h+\int_{d-h}^{d+h}\right)q(z')K(z,z')dz', \tag{1.9b}$$

where

$$K(z,z')=K_1(z,z')+K_2(z,z')=\frac{e^{-j\beta_0R_1}}{R_1}+\frac{e^{-j\beta_0R_2}}{R_2} \tag{1.9c}$$

and where

$$R_1=((z-z')^2+a^2)^{\frac{1}{2}}; \quad R_2=((z+z')^2+a^2)^{\frac{1}{2}}. \tag{1.9d}$$

The differential equation for the vector potential is the same as for a single antenna, viz.,

$$\frac{\partial^2A(z)}{\partial z^2}+\beta_0^2A(z)=j\frac{\beta_0^2}{\omega}z^iI(z); \quad (r=a). \tag{1.10a}$$

The solution consists of three equations of the form

$$A(z)=\frac{-j}{\nu_0}[C_1\cos\beta_0z+C_2\sin\beta_0z]; \quad (r=a) \tag{1.10b}$$

with different constants C_1 and C_2 in the equations applying, respectively, to the three sections of conductor extending from 0 to h, $d-h$ to d, d to $d+h$. A particular integral involving z^i as a factor has been omitted from (1.10b); it usually contributes negligibly.

Since the currents in the upper half of the central unit and the lower and upper halves of the upper unit are all different, the rigorous solution for these currents is complicated. Fortunately, an approximate solution may be obtained by making assumptions that are reasonable and that lead to integral equations like that for a single antenna.

2. THREE-ELEMENT COLLINEAR ARRAY; GENERAL ANALYSIS OF CENTRAL ANTENNA

The current $I_1(z)$ in the upper half of the central unit depends primarily upon the following integral equation valid at $r=a$ with $0\leqq z\leqq h$.

$$J_a+(J_b+J_c)=\frac{-j4\pi}{\zeta_0}[C_1\cos\beta_0z+\tfrac{1}{2}V_{10}\sin\beta_0z]. \tag{2.1}$$

For the upper sign, $V_{10}=V^s$; for the lower sign $V_{10}=V^a$. In (1),

$$J_a\equiv\int_0^h I_1(z')K(z,z')dz', \tag{2.2a}$$

$$J_b\equiv\int_{d-h}^{d}I_2(z')K(z,z')dz'$$

$$=\int_0^h I_2(d-z')K(z,d-z')dz', \tag{2.2b}$$

$$J_c\equiv\int_{d}^{d+h}I_2(z')K(z,z')dz'$$

$$=\int_0^h I_2(d+z')K(z,d+z')dz'. \tag{2.2c}$$

The current $I_1(z)$ in the central unit occurs also in two equations similar to (1) with integrals evaluated, respectively, over the halves of antenna 2. With (1) these constitute three simultaneous integral equations in the currents.

Equation (1) was obtained by substituting (1.9a) in (1.10b), evaluating $C_2=\frac{1}{2}V_{10}$, and replacing ν_0/ν_0 by its equivalent, $\zeta_0=376.7$ ohms. The last integrals in (2.2b) and (2.2c) are derived from the first integrals by changing the variable of integration to u, by setting $z'=d-u$ in (2.2b) and $z'=d+u$ in (2.2c), and then substituting z' for u. Note that each integral in (2.2a) to (2.2c) may be expanded into two integrals in the form $J_{i1}+J_{i2}$, $i=a, b, c$, by setting

$$K(z,v)=K_1(z,v)+K_2(z,v)$$

$$=\frac{e^{-j\beta_0R_1(z,v)}}{R_1(z,v)}+\frac{e^{-j\beta_0R_2(z,v)}}{R_2(z,v)}, \tag{2.3a}$$

where

$$R_1(z, v)=((z-v)^2+a^2)^{\frac{1}{2}};\quad R_2(z, v)=((z+v)^2+a^2)^{\frac{1}{2}} \quad (2.3b)$$

and

$$v=z',\quad d-z',\quad d+z'. \quad (2.3c)$$

Since the point z, where the vector potential is evaluated, is on the surface of the upper half of antenna 1 between $z=0$ and $z=h$, the contributions to the total vector potential by the integrals in (2.2a) to (2.2b) differ widely. Although the distributions of the currents in the six halves of the three antennas are not all the same, they are sufficiently alike to have the same average order of magnitude *if made equal at the centers of the antennas*. Consequently, in determining the current in the *central unit*, the currents in all units may be assumed the same. Since the currents for $z<0$ are obtained from the currents for $z>0$ using (1.8a), the following expressions relate *all* currents:

$$I_2^s(d-z')\doteq I_1^s(z')\doteq I_2^s(z'+d);\quad (0\leqq z'\leqq h) \quad (2.4a)$$

$$-I_2^a(d-z')\doteq I_1^a(z')\doteq -I_2^a(z'+d);\quad (0\leqq z'\leqq h). \quad (2.4b)$$

With (2.4a, b), (2.1) becomes

$$\int_0^h I(z')K_c(z, z')dz'$$

$$=\frac{-j4\pi}{\zeta_0}[C_1\cos\beta_0 z+\tfrac{1}{2}V_{10}\sin\beta_0 z]:$$

$$(r=a;\quad 0\leqq z\leqq h). \quad (2.5)$$

When $V_{10}=V^s$, $I=I^s$, $C_1=C_1^s$, $K_c=K_c^s$; when $V_{10}=V^a$, $I=I^a$, $C_1=C_1^a$, $K_c=K_c^a$. The new kernel, $K_c(z, z')=K_{c1}(z, z')+K_{c2}(z, z')$, is defined as follows:

$$K_c(z, z')\equiv\left\{\begin{matrix}K_c^s(z, z')\\K_c^a(z, z')\end{matrix}\right\}\equiv K(z, z')$$

$$\pm[K(z, d-z')+K(z, d+z')] \quad (2.6)$$

with $K(z, v)$ defined as in (2.3a). Note that with (2.3b)

$$R_1(z, d-z')=((z-d+z')^2+a^2)^{\frac{1}{2}}=R_1(d-z, z'), \quad (2.7a)$$

$$R_1(z, d+z')=((z-d-z')^2+a^2)^{\frac{1}{2}}=R_2(d-z, z'), \quad (2.7b)$$

$$R_2(z, d-z')=((z+d-z')^2+a^2)^{\frac{1}{2}}=R_1(d+z, z'), \quad (2.7c)$$

$$R_2(z, d+z')=((z+d+z')^2+a^2)^{\frac{1}{2}}=R_2(d+z, z'). \quad (2.7d)$$

It follows that

$$K(z, d-z')+K(z, d+z')=K(d-z, z')+K(d+z, z') \quad (2.8)$$

so that

$$K_c(z, z')=K(z, z')\pm[K(d-z, z')+K(d+z, z')]. \quad (2.9)$$

This form of the kernel is convenient in the evaluation of the expansion parameter.

Since (2.5) is an integral equation like that for the single antenna,[3a,3b] the same formal solution is obtained with a new kernel. The current in the upper half of the central unit of the collinear array is given by (3) in reference 3b with $K_c(z, z')$ substituted for $K(z, z')$. Thus,

$$I_1(z)=\frac{j2\pi V}{\zeta_0\Psi_{Kc}}\left\{\frac{\sin\beta_0(h-z)+M_{c1}(z)/\Psi_{Kc}+\cdots}{\cos\beta_0 h+A_{c1}/\Psi_{Kc}+\cdots}\right\}. \quad (2.10)$$

Similarly, the impedance of the central unit is

$$Z_{0c}=\frac{-j\zeta_0\Psi_{Kc}}{2\pi}\left\{\frac{\cos\beta_0 h+A_{c1}/\Psi_{Kc}+\cdots}{\sin\beta_0 h+B_{c1}/\Psi_{Kc}+\cdots}\right\}. \quad (2.11)$$

Each of these formulas applies to the symmetrically

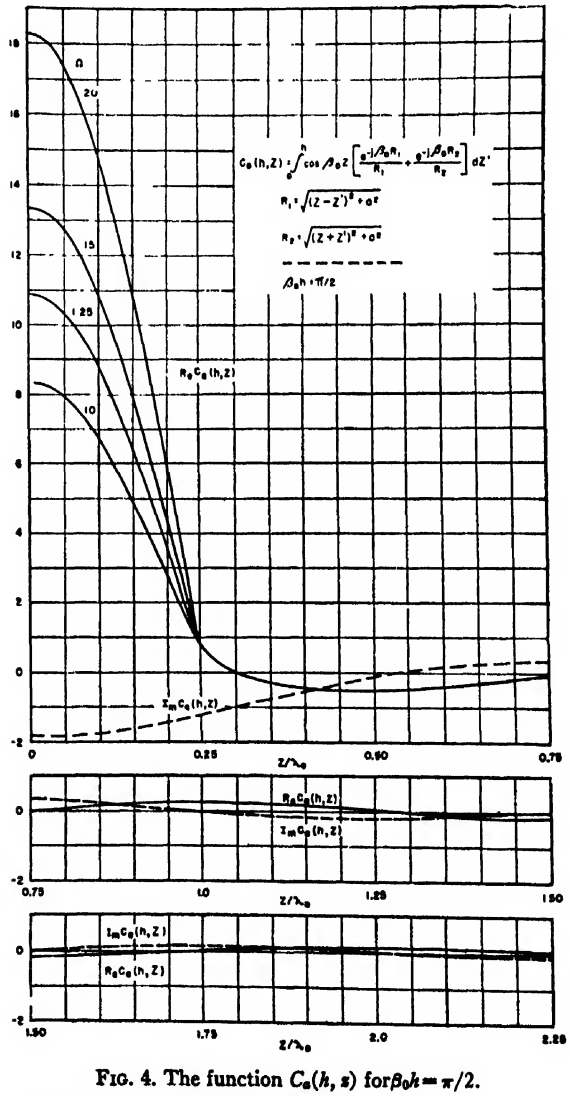


FIG. 4. The function $C_0(h, z)$ for $\beta_0 h=\pi/2$.

[3a] R. King and D. Middleton, Quart. App. Math. 3, 302–335 (January, 1946) and 4, 199–200 (July, 1946).
[3b] D. Middleton and R. King, J. App. Phys. 17, 273–284 (1946).

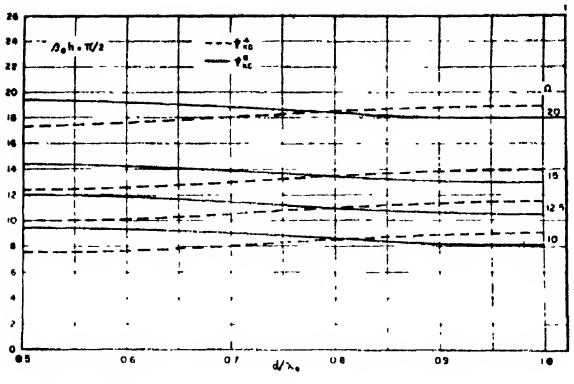


FIG. 5. The expansion parameters for the central unit in a three-element array.

driven array when a superscript s is affixed to V, I, M, B, A, Ψ and Z and the upper sign is used, and to the antisymmetrically driven array when the superscript a is affixed to these symbols and the lower sign is used. The quantities $M_{c1}(z)$, $B_{c1}=M_{1c}(0)$, and A_{c1} are the same as in reference 3 with $K_c(z, z')$ in (2.9) substituted for $K(z, z')$, and with Ψ_{Kc} replacing $\Psi=\Psi_{K1}$.

The expansion parameter Ψ_{Kc} is defined as is Ψ in reference 3. The steps include

$$\Psi_{Kc}(z)=\psi_{Kc}(z)/\sin\beta_0(h-z), \tag{2.12}$$

$$\psi_{Kc}(z)=\sin\beta_0 h\int_0^h \cos\beta_0 z' K_c(z, z')dz' - \cos\beta_0 h\int_0^h \sin\beta_0 z' K_c(z, z'^d dz'. \tag{2.13}$$

The integrals in (13) involve only the functions[4] $C_a(h, v)$ and $S_a(h, v)$ defined in references 3 and 4. Their introduction yields:

$$\psi_{Kc}(z)=\psi_1(z)\pm[\psi_1(d-z)+\psi_1(d+z)], \tag{2.14}$$

where

$$\psi_1(v)=C_a(h, v)\sin\beta_0 h-S_a(h, v)\cos\beta_0 h. \tag{2.15}$$

By far the most important case in practice is with $\beta_0 h=\pi/2$ for which

$$\psi_{Kc}(z)=C_a(h, z)\pm[C_a(h, d-z)+C_a(h, d+z)]; \quad h=\lambda_0/4. \tag{2.16}$$

The function of $C_a(h, z)$ for $\beta_0 h=\pi/2$ is given in Fig. 4. The function in (12) is:

$$\Psi_{Kc}(z)=\Psi_{K1}(z)\pm[\Psi_{K1}(d-z)+\Psi_{K1}(d+z)] \tag{2.17}$$

where $\Psi_{K1}(z)$ applies to the central unit when isolated.

Since the general behavior of $\Psi_{Kc}(z)$ as a function of z is like that of $\Psi_{K1}(z)$ for all possible values of d, the expansion parameter for the central unit may be defined as for the isolated antenna. Specifically,

$$\Psi_{Kc}=\begin{cases} |\Psi_{Kc}(0)|=|\Psi_{K1}(0)\pm 2\Psi_{K1}(d)| \\ \qquad\qquad \text{for} \quad \beta_0 h\leqq\pi/2 \\ |\Psi_{Kc}(h-\lambda_0/4)|=|\Psi_{K1}(h-\lambda_0/4) \\ \qquad \pm[\Psi_{K1}(d-h+\lambda_0/4) \\ \qquad +\Psi_{K1}(d+h-\lambda_0/4)]| \\ \qquad\qquad \text{for} \quad \beta_0 h\geqq\pi/2. \end{cases} \tag{2.18}$$

A superscript s is added to Ψ_{Kc} for $V=V^s$, and the upper signs are used; a superscript a is added for $V=-V^a$, and the lower signs are used. $\Psi_{Kc}{}^s$ and $\Psi_{Kc}{}^a$ are shown in Fig. 5.

It is clear from (2.18) that when the elements of the collinear array are driven in phase, current and impedance of the central unit are like those of an isolated antenna with a larger Ψ_{K1} and hence smaller radius. Alternatively, when the two outer units are driven 180 degrees out of phase with the central unit, this behaves like an isolated antenna with a smaller Ψ_{K1} and greater radius. This is true for $d/\lambda_0<0.8$.

The impedances Z^s and Z^a determined from (2.11) with appropriate superscripts s and a, and the corresponding currents $I_1{}^s(z)$ and $I_1{}^a(z)$ as given by (2.10) in terms of the driving voltage $V=V^s$ or $V=V^a$, are valid *only if the outer units are driven* so that the currents at their centers are equal to $I_1{}^s(0)$ in the symmetrical case, or to $-I_1{}^a(0)$ in the antisymmetrical case. Evidently, the problem of the central unit is not solved until $V_2{}^s=f^sV^s$ and $V_2{}^a=f^aV^a$ are known. They are determined in Sec. 3. Once f^s and f^a are known, a superposition of the solutions for which $V_2{}^s=V_2{}^a$ yields the solution of the array with parasitic outer elements. Since in this case

$$V^s/V^a=f^s/f^s=k, \tag{2.19a}$$

it follows that

$$\frac{1}{Z_{1\text{in}}}=Y_{1\text{in}} =\frac{I_1(0)}{V_{10}}=\frac{I_1{}^s(0)+I_1{}^a(0)}{V^s+V^a}=\frac{I_1{}^s(0)}{V^s(1+k^{-1})} +\frac{I_1{}^a(0)}{V^a(1+k)}=\frac{1}{Z_{0c}{}^s(1+k^{-1})}+\frac{1}{Z_{0c}{}^a(1+k)}. \tag{2.19b}$$

Solving for $Z_{1\text{in}}$ gives the input impedance of the three-element collinear array with the outer units parasitic.

$$Z_{1\text{in}}=\frac{2Z_{0c}{}^sZ_{0c}{}^a}{Z_{0c}{}^s+Z_{0c}{}^a}F_1=Z_{1\text{in}}{}'F_1, \tag{2.19c}$$

[4] The functions are:

$$C_a(h, z)=\int_0^h \cos\beta_0 z'\left[\frac{\exp(-j\beta_0 R_1)}{R_1}+\frac{\exp(-j\beta_0 R_2)}{R_2}\right]dz'$$

$$S_a(h, z)=\int_0^h \sin\beta_0 z'\left[\frac{\exp(-j\beta_0 R_1)}{R_1}+\frac{\exp(-j\beta_0 R_2)}{R_2}\right]dz'$$

$$R_1=((z-z')^2+a^2)^{\frac{1}{2}}; \quad R_2=((z+z')^2+a^2)^{\frac{1}{2}}.$$

where

$$F_1 \doteq 1+\frac{(Z_{0c}{}^s-Z_{0c}{}^a)(k-k^{-1})}{2Z_{0c}{}^s(1+k^{-1})+Z_{0c}{}^a(1+k)}. \quad (2.19d)$$

The determination of k is in Section 3. The numerical evaluation of Z_{11n} for $\beta_0 h=\pi/2$ is in Section 4.

The current in the central unit is given by

$$I(z)=I_1{}^s(z)+I_1{}^a(z) \quad (2.20)$$

where $I_1{}^s(z)$ is given by (2.10) with superscripts s and with $V=V^s=V_{10}/(1+k^{-1})$, and where $I_1{}^a(z)$ is given by (2.10) with superscripts a and with $V=V^a=V_{10}/(1+k)$. Once k is determined $I_1(z)$ may be calculated in terms of V_{10}, the actual driving voltage when the outer units are parasitic.

3. THREE-ELEMENT COLLINEAR ARRAY; GENERAL ANALYSIS OF OUTER ANTENNAS

The currents in the outer antennas cannot be determined so accurately as the currents in the central unit since they are not symmetrical with respect to the centers of these antennas. However, since it is not possible to center drive these units without the addition of unbalanced transmission lines, the only practically significant application of the theory as developed up to this point is with the outer units parasitic. Since in this case the currents in the parasites are not involved directly in the driving-point impedance, but only in f^s and f^a, and in obtaining the electromagnetic field of the array, it is adequate to determine approximate currents. For this purpose the principal assumption is that the current in each outer unit is composed of even and odd parts of which the odd is *not* large compared with the even. In general, this means that $\beta_0 h$ must not be near enough to π, 2π, $3\pi\cdots$ to make the odd part of the current the resonant part. When $\beta_0 h$ is near $\pi/2$ the even part of the current is resonant and the assumption is an excellent one.

The integral equation for the current in the halves of antenna 2 (Fig. 3) are

$$4\pi\nu_0 A(z_2)=\frac{-j4\pi}{\zeta_0}[C_1\cos\beta_0 z_2+C_2\sin\beta_0 z_2]; \quad (0\leqq z_2\leqq h), \quad (3.1a)$$

$$4\pi\nu_0 A(z_2)=\frac{-j4\pi}{\zeta_0}[C_1\cos\beta_0 z_2+C_4\sin\beta_0 z_2]; \quad (-h\leqq z_2\leqq 0), \quad (3.1b)$$

where

$$4\pi\nu_0 A(z_2)=\left(\int_{-h}^{h}+\int_{-d-h}^{-d+h}+\int_{-2d-h}^{-2d+h}\right)\times I(z_2')K_1(z_2, z_2')dz_2' \quad (3.2a)$$

with

$$K_1(z_2, z_2')=\frac{e^{-j\beta_0 R_1(z_2, z_2')}}{R_1(z_2, z_2')};$$

$$R_1(z_2, z_2')=((z_2-z_2')+a^2)^{\frac{1}{2}}. \quad (3.2b)$$

By setting $z'=d+z_2'$ in the second integral, $z_3'=2d+z_2'$ in the third integral and noting that

$$R_1(z_2, z'-d)=((d+z_2-z')^2+a^2)^{\frac{1}{2}}=R_1(d+z_2, z') \quad (3.3a)$$

$$R_1(z_2, z_3'-2d)=((2d+z_2-z_3')^2+a^2)^{\frac{1}{2}}=R_1(2d+z_2, z_3'). \quad (3.3b)$$

(3.2a) may be expressed as follows:

$$4\pi\nu_0 A(z_2)=\int_{-h}^{h}I_2(z_2')K_1(z_2, z_2')dz_2'+\int_{-h}^{h}I_1(z')K_1(d+z_2, z')dz'+\int_{-h}^{h}I_3(z_3')K_1(2d+z_2', z_3')dz_3'. \quad (3.4)$$

As a consequence of symmetry,

$$I_3(z_3')=I_2(-z_2'). \quad (3.5)$$

Since $I_1(z')$ is an even function of z' and symmetrical or antisymmetrical driving voltages are provided so that $I_1(z'=0)=\pm I_2(z_2'=0)$, a satisfactory approximation is

$$I_1(z')=\pm\tfrac{1}{2}[I_2(z_2')+I_2(-z_2')] \quad (3.6)$$

where the upper sign applies to the symmetrical, the lower sign to the antisymmetrical case. Substitution of (3.5) and (3.6) in (3.4) gives

$$4\pi\nu_0 A(z_2)=\int_0^h\{I_2(z_2')K_{1d}(z_2, z_2')+I_2(-z_2')K_{2d}(z_2, z_2')\}dz_2' \quad (3.7)$$

where

$$K_{1d}(z_2, z_2')\equiv K_1(z_2, z_2')\pm\tfrac{1}{2}[K_1(d+z_2, z_2')+K_2(d+z_2, z_2')]+K_2(2d+z_2, z_2') \quad (3.8a)$$

$$K_{2d}(z_2, z_2')\equiv K_2(z_2, z_2')\pm\tfrac{1}{2}[K_2(d+z_2, z_2')+K_1(d+z_2, z_2')]+K_1(2d+z_2, z_2'). \quad (3.8b)$$

The following notation is used:

$$K_1(u, z_2') \equiv \frac{e^{-j\beta_0 R_1(u, z_2')}}{R_1(u, z_2')};$$

$$K_2(u, z_2') \equiv \frac{e^{-j\beta_0 R_2(u, z_2')}}{R_2(u, z_2')}, \quad (3.9a)$$

$$R_1(u, z_2') \equiv ((u-z_2')^2+a^2)^{\frac{1}{2}};$$

$$R_2(u, z_2') = ((u+z_2')^2+a^2)^{\frac{1}{2}}. \quad (3.9b)$$

These formulas apply in the following two cases:

Symmetrical Case:

$$\left.\begin{aligned} &V_{10}=V^a; \quad V_{30}=V_{20}=V_2{}^a=f^aV^a; \\ &I_2(z_2')=I_2{}^a(z_2') \\ &K_{1d}(z_2, z_2') \text{ with upper sign} = K_{1d}{}^a(z_2, z_2'). \end{aligned}\right\}, \quad (3.10a)$$

Antisymmetrical Case:

$$\left.\begin{aligned} &V_{10}=V^a; \quad V_{30}=V_{20}=-V_2{}^a=-f^aV^a; \\ &I_2(z_2')=I_2{}^a(z_2') \\ &K_{1d}(z_2, z_2') \text{ with lower sign} = K_{1d}{}^a(z_2, z_2') \end{aligned}\right\}. \quad (3.10b)$$

The first step in the solution of (3.1a) with (3.7) is to resolve both sides into even and odd functions of z_2 and in this way obtain two equations, the one for the even part of $I_2(z_2)$, the other for the odd part. The conditions for the even parts of the vector potential and current are

$$A_{\text{even}}(z_2) = \tfrac{1}{2}[A(z_2)+A(-z_2)];$$

$$I_{\text{even}}(z_2) = \tfrac{1}{2}[I_2(z_2)+I_2(-z_2)] \quad (3.11a)$$

so that with (3.1a)

$$4\pi\nu_0 A_{\text{even}}(z_2) = \frac{-j4\pi}{\zeta_0}[C_{\text{even}} \cos\beta_0 z_2 + \tfrac{1}{2}V_{20} \sin\beta_0 z_2];$$

$$-h \leqq z_2 \leqq h) \quad (3.11b)$$

where $C_{\text{even}} = \frac{1}{2}(C_1+C_3)$ and $\frac{1}{2}V_{20} = \frac{1}{2}(C_2-C_4)$. The even part of the vector potential in the range $(-h \leqq z_2 \leqq 0)$ is obtained from $A_{\text{even}}(-z_2) = A_{\text{even}}(z_2)$.

The conditions for the odd part of the vector potential and current are

$$A_{\text{odd}}(z_2) = \tfrac{1}{2}[A(z_2)-A(-z_2)];$$

$$I_{\text{odd}}(z_2) = \tfrac{1}{2}[I_2(z_2)-I_2(-z_2)] \quad (3.12a)$$

so that with (1a)

$$4\pi\nu_0 A_{\text{odd}}(z_2) = \frac{-j4\pi}{\zeta_0} C_{\text{odd}} \sin\beta_0 z_2; \quad (-h \leqq z_2 \leqq h) \quad (3.12b)$$

where $C_{\text{odd}} = \frac{1}{2}(C_3+C_4)$ and $\frac{1}{2}(C_1-C_3)=0$ since $A_{\text{odd}}(-z_2) = A_{\text{odd}}(z_2) = 0$ at $z_2=0$. The odd vector potential in the range $(-h \leqq z_2 \leqq 0)$ is obtained from $A_{\text{odd}}(-z_2) = -A_{\text{odd}}(z_2)$.

The integrals and equations for the even and odd parts of the vector potential are obtained using (3.11a, b) and (3.12a, b). They are

$$4\pi\nu_0 A_{\text{even}}(z_2) = \int_0^h \{I_2(z_2')K_{1\text{even}}(z_2, z_2') + I_2(-z_2')K_{2\text{even}}(z_2, z_2')\}dz_2'$$

$$= \frac{-j4\pi}{\zeta_0}[C_{\text{even}} \cos\beta_0 z_2 + \tfrac{1}{2}V_{20} \sin\beta_0 |z_2|];$$

$$(-h \leqq z_2 \leqq h) \quad (3.13a')$$

$$4\pi\nu_0 A_{\text{odd}}(z_2) = \int_0^h \{I_2(z_2')K_{1\text{odd}}(z_2, z_2') + I_2(-z_2')K_{2\text{odd}}(z_2, z_2')\}dz_2$$

$$= \frac{-j4\pi}{\zeta_0} C_{\text{odd}} \sin\beta_0 z_2; \quad (-h \leqq z_2 \leqq h) \quad (3.13b)$$

where

$$K_{1\text{even}}(z_2, z_2') \equiv \tfrac{1}{2}[K_{1d}(z_2, z_2') + K_{1d}(-z_2, z_2')], \quad (3.14a)$$

$$K_{2\text{even}}(z_2, z_2') \equiv \tfrac{1}{2}[K_{2d}(z_2, z_2') + K_{2d}(-z_2, z_2')], \quad (3.14b)$$

$$K_{1\text{odd}}(z_2, z_2') \equiv \tfrac{1}{2}[K_{1d}(z_2, z_2') - K_{1d}(-z_2, z_2')], \quad (3.14c)$$

$$K_{2\text{odd}}(z_2, z_2') = \tfrac{1}{2}[K_{2d}(z_2, z_2') - K_{2d}(-z_2, z_2')]. \quad (3.14d)$$

In an asymmetrical geometrical arrangement like that of the outer elements in a collinear array, contributions to the even as well as the odd parts of the vector potential come from *both* even and odd parts of the currents in the three units. Consequently (3.13a) and (3.13b) cannot be solved directly for $I_{\text{even}}(z_2)$ and $I_{\text{odd}}(z_2)$. However, an approximate procedure is available. The current may be expanded into its even and odd parts by setting

$$I_2(z_2') = I_{\text{even}}(z_2') + I_{\text{odd}}(z_2');$$

$$I_2(-z_2') = I_{\text{even}}(z_2') - I_{\text{odd}}(z_2') \quad (3.15a)$$

where by definition

$$I_{\text{even}}(z_2') = \tfrac{1}{2}[I(z_2') + I_2(-z_2')] = I_{\text{even}}(-z_2'), \quad (3.15b)$$

$$I_{\text{odd}}(z_2') = \tfrac{1}{2}[I(z_2') - I_2(-z_2')] = -I_{\text{odd}}(-z_2'). \quad (3.15c)$$

If (3.15a) is substituted in (3.13a) and (3.13b), the

results are:

$$4\pi\nu_0 A_{\text{even}}(z_2)=\int_0^h \{I_{\text{even}}(z_2')[K_{1\text{even}}(z_2, z_2') + K_{2\text{even}}(z_2, z_2')]+I_{\text{odd}}(z_2')[K_{1\text{even}}(z_2, z_2') - K_{2\text{even}}(z_2, z_2')]\}dz_2' \quad (3.16a)$$

$$4\pi\nu_0 A_{\text{odd}}(z_2)=\int_0^h \{I_{\text{even}}(z_2')[K_{1\text{odd}}(z_2, z_2') + K_{2\text{odd}}(z_2, z_2')]+I_{\text{odd}}(z_2')[K_{1\text{odd}}(z_2, z_2') - K_{2\text{odd}}(z_2, z_2')]\}dz_2'. \quad (3.16b)$$

Let the following kernels be defined:

$$K_{\text{even}}(z_2, z_2')\equiv K_{1\text{even}}(z_2, z_2')+K_{2\text{even}}(z_2, z_2')=\tfrac{1}{2}[K_{1d}(z_2, z_2')+K_{2d}(z_2, z_2')+K_{1d}(-z_2, z_2')+K_{2d}(-z_2, z_2')]$$
$$=K_1(z_2, z_2')+K_2(z_2, z_2')\pm\tfrac{1}{2}[K_1(d+z_2, z_2')+K_2(d+z_2, z_2')+K_1(d-z_2, z_2')+K_2(d-z_2, z_2')]$$
$$+\tfrac{1}{2}[K_1(2d+z_2, z_2')+K_2(2d+z_2, z_2')+K_1(2d-z_2, z_2')+K_2(2d-z_2, z_2')] \quad (3.17a)$$

$$K_{e0}(z_2, z_2')\equiv K_{1\text{even}}(z_2, z_2')-K_{2\text{even}}(z_2, z_2')=\tfrac{1}{2}[K_{1d}(z_2, z_2')-K_{2d}(z_2, z_2')+K_{1d}(-z_2, z_2')-K_{2d}(-z_2, z_2')]$$
$$=-\tfrac{1}{2}[K_1(2d+z_2, z_2')-K_2(2d+z_2, z_2')+K_1(2d-z_2, z_2')-K_2(2d-z_2, z_2')] \quad (3.17b)$$

$$K_{0e}(z_2, z_2')\equiv K_{1\text{odd}}(z_2, z_2')+K_{2\text{odd}}(z_2, z_2')=\tfrac{1}{2}[K_{1d}(z_2, z_2')+K_{2d}(z_2, z_2')-K_{1d}(-z_2, z_2')-K_{2d}(z_2, z_2')]$$
$$=\pm\tfrac{1}{2}[K_1(d+z_2, z_2')+K_2(d+z_2, z_2')-K_1(d-z_2, z_2')-K_2(d-z_2, z_2')]$$
$$+\tfrac{1}{2}[K_1(2d+z_2, z_2')+K_2(2d+z_2, z_2')-K_1(2d-z_2, z_2')-K_2(2d-z_2, z_2')] \quad (3.17c)$$

$$K_{\text{odd}}(z_2, z_2')\equiv K_{1\text{odd}}(z_2, z_2')-K_{2\text{odd}}(z_2, z_2')=\tfrac{1}{2}[K_{1d}(z_2, z_2')-K_{2d}(z_2, z_2')-K_{1d}(-z_2, z_2')+K_{2d}(-z_2, z_2')]$$
$$=K_1(z_2, z_2')-K_2(z_2, z_2')-\tfrac{1}{2}[K_1(2d+z_2, z_2')-K_2(2d+z_2, z_2')]$$
$$+\tfrac{1}{2}[K_1(2d-z_2, z_2')-K_2(2d-z_2, z_2')]. \quad (3.17d)$$

Examination of (3.17b) as it appears in (3.16a) shows that it represents the contribution to the even part of the vector potential on antenna 2 by the odd parts of the currents in antenna 3. Since the odd currents involved are assumed to be relatively small, and in any case the distances are large, this term is negligible compared with the very much greater contribution from the large even currents in all three antennas. Since $R_1(2d\pm z_2, z_2')$ differs very much less from $R_2(2d\pm z_2, z_2')$ than does $R_1(z_2, z_2')$ from $R_2(z_2, z_2')$ it follows that the principal contributions to $K_{\text{odd}}(z_2, z_2')$ in (3.17d) are from the first two terms. That is,

$$K_{\text{odd}}(z_2, z_2')\doteq K_1(z_2, z_2')-K_2(z_2, z_2'). \quad (3.17e)$$

Neglecting the term in $K_{e0}(z_2, z_2')$, the two integral equations are:

$$4\pi\nu_0 A_{\text{even}}(z_2)=\int_0^h I_{\text{even}}(z_2')K_{\text{even}}(z_2, z_2')dz_2$$
$$=\frac{-j4\pi}{\zeta_0}[C_{\text{even}}\cos\beta_0 z_2+\tfrac{1}{2}V_{20}\sin\beta_0 z_2]; \quad (-h\leqq z_2\leqq h) \quad (3.18a)$$

$$4\pi\nu_0 A_{\text{odd}}(z_2)=\int_0^h I_{\text{even}}(z_2')K_{0e}(z_2, z_2')dz_2'$$
$$+\int_0^h I_{\text{odd}}(z_2')K_{\text{odd}}(z_2, z_2')dz_2'$$
$$=\frac{-j4\pi}{\zeta_0}C_{\text{odd}}\sin\beta_0 z_2; \quad (-h\leqq z_2\leqq h). \quad (3.18b)$$

Of these the first may be solved directly for $I_{\text{even}}(z_2)$; using this known value, the second equation may be solved for $I_{\text{odd}}(z_2)$. Note that the equations apply only in the symmetrical and antisymmetrical cases defined in (3.10a, b) but that the general case may be constructed by superposition.

The Eq. (3.18a) for the even part of the current is in exactly the same form as the equation for a single antenna. The solution may be written down directly in the form

$$I_{\text{even}}(z_2)=\frac{j2\pi V_{20}}{\zeta_0\Psi_{\text{even}}}\times\left\{\frac{\sin\beta_0(h-z_2)+M_{\text{even1}}(z_2)/\Psi_{\text{even}}+\cdots}{\cos\beta_0 h+A_{\text{even1}}/\Psi_{\text{even}}+\cdots}\right\}. \quad (3.19a)$$

The impedance is

$$Z_{\text{even}}=\frac{-j\zeta_0\Psi_{\text{even}}}{2\pi}\left\{\frac{\cos\beta_0 h+A_{\text{even1}}/\Psi_{\text{even}}+\cdots}{\sin\beta_0 h+B_{\text{even1}}/\Psi_{\text{even}}+\cdots}\right\}. \quad (3.19b)$$

For the symmetrical case a superscript s is added to I, Ψ, M, B, A, and Z, and V_{20} is replaced by V_2^s; for the antisymmetrical case a superscript a replaces the superscript s. The functions M, B, A, and Ψ are defined as for a single antenna but using a different kernel.

An important object in evaluating $I_{\text{even}}^s(z_2)$ and $I_{\text{odd}}^a(z_2)$ is to determine the factor f^s and f^a. They may be evaluated directly from (3.19a, b) as follows. Since the odd part of $I_2(z_2)$ necessarily is zero at $z_2=0$, the

entire current at this point is even. Hence, using (1.6) and (1.7),

Symmetrical case:

$$V_{20}=V_2{}^s=I_2{}^s(z_2=0)Z_{\text{even}}{}^s$$

$$=I_1{}^s(0)Z_{\text{even}}{}^s=\frac{V^s}{Z_{0c}{}^s}\cdot Z_{\text{even}}{}^s=f^sV^s. \quad (3.20a)$$

Therefore,

$$f^s=Z_{\text{even}}{}^s/Z_{0c}{}^s. \quad (3.20b)$$

Antisymmetrical case:

$$V_{20}=-V_2{}^a=I_2{}^a(z_2=0)Z_{\text{even}}{}^a$$

$$=-I_1{}^a(0)Z_{\text{even}}{}^a=\frac{-V^a}{Z_{0c}{}^a}Z_{\text{even}}{}^a=-f^aV^a. \quad (3.21a)$$

Therefore,

$$f^a=Z_{\text{even}}{}^a/Z_{0c}{}^a. \quad (3.21b)$$

$Z_{0c}{}^s$ and $Z_{0c}{}^a$ are given by (2.11) with appropriate superscripts. The complex factor k defined in (2.19a) is given by

$$k=f^a/f^s=\frac{Z_{\text{even}}{}^a}{Z_{\text{even}}{}^s}\cdot\frac{Z_{0c}{}^s}{Z_{0c}{}^a}. \quad (3.22)$$

It is this factor that is required in order to calculate the input impedance of the three-element array with the outer units parasitic using (2.19c).

The even part of the current in the outer antennas when these are parasitic is given by

$$I_{2\text{even}}(z_2)=I_{\text{even}}{}^s(z_2)-I_{\text{even}}{}^a(z_2) \quad (3.23a)$$

where $I_{\text{even}}{}^s(z_2)$ is obtained from (3.19a) with appropriate superscripts s and with $V_{20}=V_2{}^s$, and where $I_{\text{even}}{}^a(z_2)$ is given by (3.19a) with superscripts a and with $V_{20}=V_2{}^a$. Note that

$$V_2{}^a=V_2{}^s=f^sV^s=f^aV^a=V_{10}/2F_2 \quad (3.23b)$$

where

$$F_2\equiv(f^a+f^s)/2f^af^s. \quad (3.23c)$$

Thus,

$$I_{2\text{even}}(z_2)=\frac{j\pi V_{10}}{\zeta_0F_2}\left\{\frac{1}{\Psi_{\text{even}}{}^s}\left[\frac{\sin\beta_0(h-z_2)+M_{\text{even}1}{}^s(z_2)/\Psi_{\text{even}}{}^s+\cdots}{\cos\beta_0h+A_{\text{even}1}{}^s/\Psi_{\text{even}}{}^s+\cdots}\right]\right.$$

$$\left.-\frac{1}{\Psi_{\text{even}}{}^a}\left[\frac{\sin\beta_0(h-z_2)+M_{\text{even}1}{}^a(z_2)/\Psi_{\text{even}}{}^a+\cdots}{\cos\beta_0h+A_{\text{even}1}{}^a/\Psi_{\text{even}}{}^a+\cdots}\right]\right\}. \quad (3.23d)$$

$I_{2\text{even}}(z_2)$ does not differ greatly from $I_1(z)$ since both are even functions that are equal at the centers of all three antennas.

The odd part of the current in antenna 2 must be determined from (3.18b) using the known value of $I_{\text{even}}(z_2)$ obtained from the solution (3.19a) of (3.18a). Thus,

$$\int_0^h I_{\text{odd}}(z_2')K_{\text{odd}}(z_2,z_2')dz_2'$$

$$=\frac{-j4\pi}{\zeta_0}[C_{\text{odd}}\sin\beta_0z_2+U(z_2)];\quad(0\leqq z_2\leqq h) \quad (3.24a)$$

where

$$U(z_2)\equiv\frac{-j\zeta_0}{4\pi}\int_0^h I_{\text{even}}(z_2')K_{0e}(z_2,z_2')dz_2'. \quad (3.24b)$$

$K_{\text{odd}}(z_2,z_2')$ is given by (3.17e) and $K_{0e}(z_2,z_2')$ by (3.17c). The upper sign in (3.17c) applies to the symmetrical, the lower sign to the antisymmetrical case. Note that $K_{\text{odd}}(z_2,z_2')$ is the same for both. The evaluation of $I_{\text{odd}}(z_2)$ from (3.24a) is involved particularly when β_0h is near π since it then is a resonant-mode current whereas the even current is not. On the other hand, when β_0h is near $\pi/2$ the even distribution is resonant, the odd is not and, therefore, relatively small. Since the only practically important case, and at the same time the one for which an approximate evaluation of (3.24a) is readily carried out, is with $\beta_0h=\pi/2$, only this is considered.

The first step in the solution of (3.24a) is the simplification of the function $U(z_2)$. Since z_2 cannot exceed h, and $d\geqq 2h$ occurs in all distances R in $K_{0e}(z_2,z_2')$, the term a^2 in these distances is negligible. It follows that in the range $0\leqq z_2\leqq h$, $0\leqq z_2'\leqq h$, (3.17c) may be simplified into

$$K_{0e}(z_2,z_2')\doteq\pm\tfrac{1}{2}e^{-j\beta_0(d+z_2)}\left[\frac{e^{j\beta_0z_2'}}{d+z_2-z_2'}+\frac{e^{-j\beta_0z_2'}}{d+z_2+z_2'}\right]$$

$$\mp\tfrac{1}{2}e^{-j\beta_0(d-z_2)}\left[\frac{e^{j\beta_0z_2'}}{d-z_2-z_2'}+\frac{e^{-j\beta_0z_2'}}{d-z_2+z_2'}\right]$$

$$+\tfrac{1}{2}e^{-j\beta_0(3d+z_2)}\left[\frac{e^{j\beta_0z_2'}}{2d+z_2-z_2'}+\frac{e^{-j\beta_0z_2'}}{2d+z_2+z_2'}\right]$$

$$+\tfrac{1}{2}e^{-j\beta_0(2d-z_2)}\left[\frac{e^{j\beta_0z_2'}}{2d-z_2-z_2'}+\frac{e^{-j\beta_0z_2'}}{2d-z_2+z_2'}\right]. \quad (3.25)$$

As a next step in the simplification let z_2' be neglected in the denominators throughout. This is a good approximation for all values of z_2' in most of the terms, a poor approximation only in one term and then only when

$z_2=h$ and when z_2' approaches h. Since the current vanishes at h, an error here is of no great importance. Hence, setting $z_2'=0$ in the denominators, (3.25) reduces to

$$K_{0e}(z_2, z_2')=\left[\pm\frac{e^{-j\beta_0(d+z_2)}}{d+z_2}\mp\frac{e^{-j\beta_0(d-z_2)}}{d-z_2}+\frac{e^{-j\beta_0(2d+z_2)}}{2d+z_2}+\frac{e^{-j\beta_0(2d-z_2)}}{2d-z_2}\right]\cos\beta_0 z_2'. \quad (3.26)$$

Since with $\beta_0 h=\pi/2$, the even current is well approximated by

$$I_{\text{even}}(z_2)\equiv\frac{V}{Z_{0e}}\cos\beta_0 z_2' \quad (3.27)$$

the integral in (3.24b) may be evaluated directly using (3.26). It involves

$$\int_0^h \cos^2\beta_0 z_2' dz_2'=\pi/4\beta_0. \quad (3.28)$$

Accordingly,

$$U(z_2)\doteq\frac{j\zeta_0 V}{8Z_{0e}}\left\{\frac{\pm e^{-j\beta_0 d}}{\beta_0(d^2-z_2^2)}[z_2\cos\beta_0 z_2+jd\sin\beta_0 z_2]+\frac{e^{-j2\beta_0 d}}{\beta_0(4d^2-z^2)}[z_2\cos\beta_0 z_2+j2d\sin\beta_0 z_2]\right\}. \quad (3.29)$$

In the range $0\leqq\beta_0 z_2\leqq\pi/2$, the approximation $\beta_0 z_2\doteq\sin\beta_0 z_2$ is satisfactory except near $\beta_0 z_2=\pi/2$ where, however, the current vanishes and an error in $U(z_2)$ is immaterial. With this approximation,

$$U(z_2)\doteq\frac{\zeta_0 V}{8Z_{0e}\beta_0 d}F(z_2)\sin\beta_0 z_2\equiv\bar{U}(z_2)\sin\beta_0 z_2 \quad (3.30a)$$

where $\bar{U}(z_2)$ is defined by (3.30a) and where

$$F(z_2)=-\left\{\frac{\pm e^{-j\beta_0 d}}{(1-z_2^2/d^2)}\left[1-j\frac{\cos\beta_0 z_2}{\beta_0 d}\right]+\frac{e^{-j2\beta_0 d}}{(1-z_2^2/4d^2)}\left[1-j\frac{\cos\beta_0 z_2}{2\beta_0 d}\right]\right\}. \quad (3.30b)$$

Since $z_2/d\leqq\frac{1}{4}$, and $\beta_0 d\geqq\pi$, $F(z_2)$ does not vary rapidly. It follows that $U(z_2)$ varies approximately as $\sin\beta_0 z_2$. With (3.30a, b) the integral equation (3.24a) reduces to

$$\int_0^h I_{\text{odd}}(z_2')K_{\text{odd}}(z_2, z_2')dz_2'=\frac{-j4\pi}{\zeta_0}[C_{\text{odd}}+\bar{U}(z_2)]\sin\beta_0 z_2. \quad (3.31)$$

This integral equation can be solved for the current by

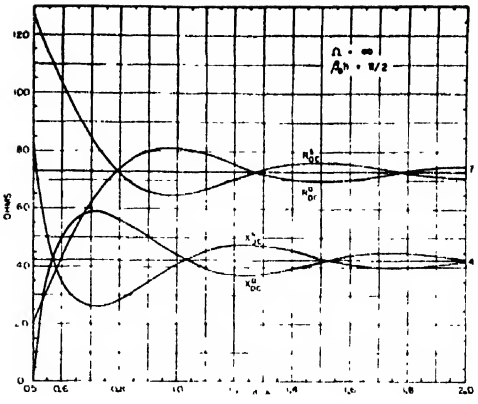


FIG. 6. Symmetrical and antisymmetrical impedances of the central unit in a three-antenna array.

the usual method of iteration. Let

$$I_{\text{odd}}(z_2')=I_{\text{odd}}(z_2)g(z_2, z_2') \quad (3.32)$$

where $g(z_2, z_2')$ is a distribution function that is chosen to make $I_{\text{odd}}(z_2)\Psi_{\text{odd}}(z_2)$ approximate closely the integral on the left in (3.31) with

$$\Psi_{\text{odd}}(z_2)=\int_0^h g(z_2, z_2')K_{\text{odd}}(z_2, z_2')dz_2'=\Psi_{\text{odd}}+\gamma_{\text{odd}}(z_2). \quad (3.33)$$

Ψ_{odd} is a mean constant value of the integral defined at a suitable reference point. With (3.32) and (3.33)

$$I_{\text{odd}}(z_2)=\frac{-j4\pi}{\zeta_0\Psi_{\text{odd}}}[C_{\text{odd}}+\bar{U}(z_2)]\sin\beta_0 z_2-\frac{1}{\Psi_{\text{odd}}}\left[I_{\text{odd}}(z_2)\gamma_{\text{odd}}(z_2)+\int_0^h[I_{\text{odd}}(z_2')-I_{\text{odd}}(z_2)g(z_2, z_2')]K_{\text{odd}}(z_2, z_2')dz_2'\right]. \quad (3.34)$$

By using the term in $\sin\beta_0 z_2$ as the zeroth-order current, and substituting this in the integrals, these may be evaluated assuming $\bar{U}(z_2)$ to be constant, and added to the zeroth-order current. This process of iteration can be repeated. The current then appears in the form

$$I_{\text{odd}}(z_2)=\frac{-j4\pi}{\zeta_0\Psi_{\text{odd}}}[C_{\text{odd}}+\bar{U}(z_2)]\times\left[\sin\beta_0 z_2+\frac{S_1(z_2)}{\Psi_{\text{odd}}}+\frac{S_2(z_2)}{\Psi_{\text{odd}}^2}+\cdots\right]. \quad (3.35)$$

The constant C_{odd} may be evaluated by requiring the current to vanish at $z_2=h$. Then

$$C_{\text{odd}}=-U(h) \quad (3.36)$$

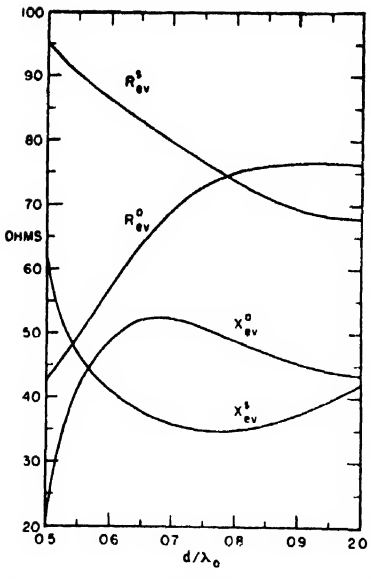


FIG. 7. Symmetrical and antisymmetrical impedances of the outer units in a three-element array.

so that

$$I_{\text{odd}}(z_2)=\frac{-j4\pi}{\zeta_0\Psi_{\text{odd}}}[U(z_2)-U(h)]\times\left[\sin\beta_0 z_2+\frac{S_1(z_2)}{\Psi_{\text{odd}}}+\frac{S_2(z_2)}{\Psi_{\text{odd}}^2}+\cdots\right]. \quad (3.37)$$

It follows from (3.30b) that with $\beta_0 h=\pi/2$, an approximate result is:

$$U(z_2)-U(h)\doteq\frac{j\zeta_0 VD(d)}{8Z_{0c}}\cos\beta_0 z_2, \quad (3.38a)$$

where

$$D(d)\equiv\frac{1}{\beta_0^2 d^2}\left[\pm\frac{e^{-j\beta_0 d}}{(1-h^2/d^2)}+\frac{e^{-j2\beta_0 d}}{2(1-h^2/4d^2)}\right]. \quad (3.38b)$$

Substitution of (3.38a) in (3.37) gives the following simple expressions for the zeroth-order odd currents in the outer units when $\beta_0 h=\pi/2$:

$$I_{\text{odd}^s}(z_2)=\frac{V^s}{Z_{0c^s}}\frac{\pi D^s(d)}{4\Psi_{\text{odd}}}\sin 2\beta_0 z_2 \quad (3.39a)$$

$$I_{\text{odd}^a}(z_2)=\frac{V^a}{Z_{0c^a}}\frac{\pi D^a(d)}{4\Psi_{\text{odd}}}\sin 2\beta_0 z_2 \quad (3.39b)$$

where $D^s(d)$ is (3.38b) with the upper sign and $D^a(d)$ is (3.38b) with the lower sign. Note that $V^s/Z_{0c^s}=I_1^s(0)$; $V^a/Z_{0c^a}=I_1^a(0)$. The odd part of the current in the outer units when these are parasitic is given by the sum of (3.39a) and (3.39b). Noting that as in (2.19b), $V^s=V_{10}/(1+k^{-1})$, $V^a=V_{10}/(1+k)$, it follows that

$$I_{2\text{odd}}(z_2)=I_{\text{odd}^s}(z_2)+I_{\text{odd}^a}(z_2)=\frac{V_{10}\pi}{4\Psi_{\text{odd}}}\left[\frac{D^s(d)}{Z_{0c^s}(1+k^{-1})}+\frac{D^a(d)}{Z_{0c^a}(1+k)}\right]\sin 2\beta_0 z_2;$$

$$(\beta_0 h=\pi/2). \quad (3.40)$$

The total current in the outer antennas is the sum of (3.40) and (3.23d) with $\beta_0 h=\pi/2$. Note that as the antenna is made thinner, i.e. $\Omega=2\ln(2h/a)$ increased, the odd current does not approach a value independent of Ω when $\beta_0 h=\pi/2$ as does the even current.

The final step in the complete evaluation of $I_{\text{odd}}(z_2)$ is the determination of Ψ_{odd} from (3.33). Since the zeroth-order distribution leads to a distribution function $g(z_2, z_2')=\sin 2\beta_0 z'/\sin 2\beta_0 z$ that does not permit the ready evaluation of Ψ_{odd}, the simpler procedure of assuming a sensibly uniform current and neglecting retardation is more convenient and adequate. This is equivalent to setting

$$\Psi_{\text{odd}}(z_2)=\int_0^h\left[\frac{1}{[(z_2'-z_2)^2+a^2]^{\frac{1}{2}}}-\frac{1}{[(z_2'+z_2)^2+a^2]^{\frac{1}{2}}}\right]dz_2'. \quad (3.41)$$

Since the odd current is most nearly constant near $z_2=h/2$ and this is its greatest value, it is a good approximation to choose the reference point for Ψ_{odd} at $z_2=h/2$. Thus

$$\Psi_{\text{odd}}=\int_{-h/2}^{h/2}\frac{du}{(u^2+a^2)^{\frac{1}{2}}}-\int_{h/2}^{3h/2}\frac{du}{(u^2+a^2)^{\frac{1}{2}}}=\ln\left(\frac{h^2}{3a^2}\right)=\Omega-2.5, \quad (3.42)$$

where $\Omega=2\ln(2h/a)$.

In this section general expressions have been obtained for the symmetrical and antisymmetrical components of the even current in the outer units of the three-element array. An approximate formula for the odd currents has been derived in the important special case, $\beta_0 h=\pi/2$. As a part of the determination of the even currents, relations between the three driving voltages were established which assure that the currents at the centers of all three units are equal in magnitude and either in-phase or 180 degrees out-of-phase as required by the conditions for the symmetrical and antisymmetrical cases. Thus, in Section 2 and this section the general formulas for the currents in collinear arrays are derived in the special case of the three-element array.

4. COLLINEAR ARRAY OF THREE HALF-WAVE DIPOLES

The most useful collinear array of three antennas consists of identical units each of electrical half-length $\beta_0 h = \pi/2$. The middle antenna No. 1 is center-driven from a balanced two-wire transmission line. As with the single antenna[1] this line may be replaced by an appropriate terminal-zone network with a slice generator. The two outer antennas, No. 2 and No. 3, are parasitic as shown in Fig. 2b.

The driving voltages for the three units are

$$V_{10} = V^s + V^a = V^s(1+k^{-1}) = V^a(1+k);$$

$$V_{20} = V_{30} = V_2{}^s - V_2{}^a = f^s V^s - f^a V^a = 0 \quad (4.1a)$$

where

$$k = \frac{f^a}{f^s}; \quad f^a = \frac{Z_{\text{even}}{}^a}{Z_{0c}{}^a}; \quad f^s = \frac{Z_{\text{even}}{}^s}{Z_{0c}{}^s}, \quad (4.1b)$$

and where Z_{even} is the impedance of each outer unit, Z_{0c} of the central unit. V^s is the driving voltage of the central unit when the other two are center-driven by voltages $V_3{}^s = V_2{}^s = f^s V^s$ such that the *currents* at the *centers* of all three units are equal and in phase. V^a is the driving voltage of the central unit when the two outer antennas are center-driven by voltages $V_3{}^a = V_2{}^a = f^a V^a$ such that the currents at the centers of the two outer units are equal to and in-phase with each other, but both equal to and 180 degrees out-of-phase with the current at $z=0$ in the central unit.

The first-order symmetrical and antisymmetrical currents in the central unit are given by (2.10) with $\beta_0 h = \pi/2$ and appropriate superscripts. Thus,

$$I_1{}^s(z) = \frac{j2\pi V^s}{\zeta_0 A_{c1}{}^s}[\cos\beta_0 z + M_{c1}{}^s(z)/\Psi_{Kc}{}^s];$$

$$(0 \leqq z \leqq h) \quad (4.2a)$$

$$I_1{}^a(z) = \frac{j2\pi V^a}{\zeta_0 A_{c1}{}^a}[\cos\beta_0 z + M_{c1}{}^a(z)/\Psi_{Kc}{}^a];$$

$$(0 \leqq z \leqq h). \quad (4.2b)$$

The total first-order current in the central unit when the outer units are parasitic is the sum of (4.2a) and (4.2b). With (4.1a, b) it is,

$$I_1(z) = I_1{}^s(z) + I_1{}^a(z) = \frac{j2\pi V_{10}}{\zeta_0}$$

$$\times\left\{\left[\frac{1}{A_{c1}{}^s(1+k^{-1})} + \frac{1}{A_{c1}{}^a(1+k)}\right]\cos\beta_0 z\right.$$

$$\left.+\left[\frac{M_{c1}{}^s(z)}{A_{c1}{}^s(1+k^{-1})\Psi_{K1}{}^s} + \frac{M_{c1}{}^a(z)}{A_{c1}{}^a(1+k)\Psi_{K1}{}^a}\right]\right\}. \quad (4.3)$$

The first-order symmetrical and antisymmetrical impedances of the central unit are

$$Z_{0c}{}^s = \frac{-j\zeta_0 A_{c1}{}^s}{2\pi}[1 + B_{c1}{}^s/\Psi_{Kc}{}^s]^{-1};$$

$$Z_{0c}{}^a = \frac{-j\zeta_0 A_{c1}{}^a}{2\pi}[1 + B_{c1}{}^a/\Psi_{Kc}{}^a]^{-1}. \quad (4.4a)$$

The first-order input impedance of the array when the outer units are parasitic is

$$Z_{1\text{in}} = \frac{-j\zeta_0}{2\pi}\left\{\left[\frac{1}{A_{c1}{}^s(1+k^{-1})} + \frac{1}{A_{c1}{}^a(1+k)}\right]\right.$$

$$\left.+\left[\frac{B_{c1}{}^s}{A_{c1}{}^s(1+k^{-1})\Psi_{K1}{}^s} + \frac{B_{c1}{}^a}{A_{c1}{}^a(1+k)\Psi_{K1}{}^a}\right]\right\}^{-1}. \quad (4.4b)$$

where

$$B_{c1}{}^s \equiv M_{c1}{}^s(0); \quad B_{c1}{}^a \equiv M_{c1}{}^a(0). \quad (4.4c)$$

For antennas that have a reasonably large value of h/a, $\Psi_{K1}{}^s$ and $\Psi_{K1}{}^a$ are sufficiently great to make the contributions by the $1/\Psi_{K1}$ terms in (2.2a, b), (2.3) and (2.4a, b) relatively small. When this is true the following expressions are satisfactory approximations:

$$I_1{}^s(z) \doteq \frac{V^s}{Z_{0c}{}^s}\cos\beta_0 z; \quad I_1{}^a(z) \doteq \frac{V^a}{Z_{0c}{}^a}\cos\beta_0 z;$$

$$I_1(z) = \frac{V_{10}}{Z_{1\text{in}}}\cos\beta_0 z \quad (4.5a)$$

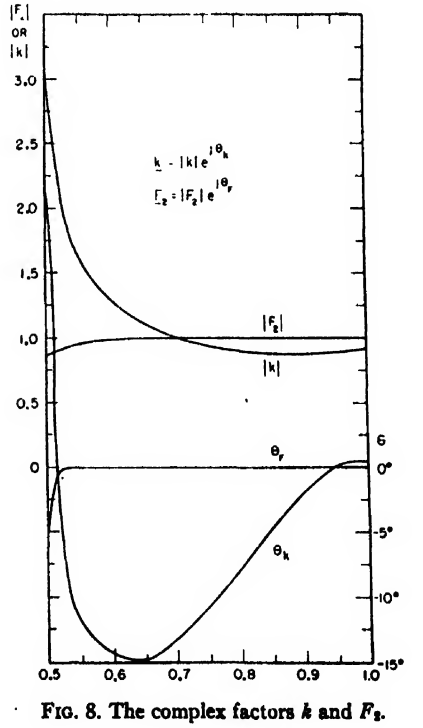


FIG. 8. The complex factors k and F_z.

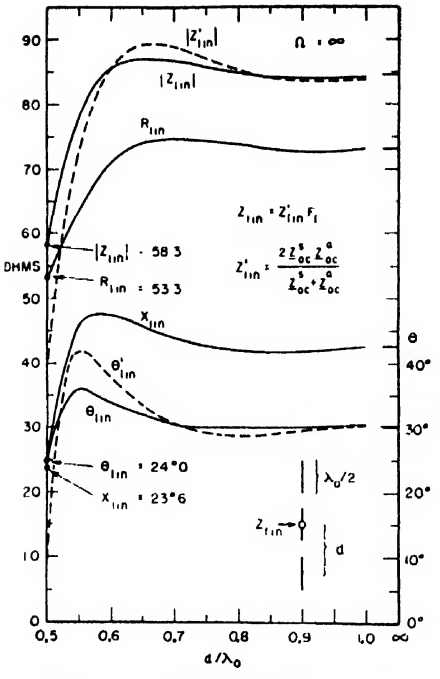


FIG. 9. Input impedance of antenna with two collinear parasites.

where

$$Z_{0c}{}^{s} \doteq \frac{-j\zeta_0 A_{c1}{}^{s}}{2\pi}; \quad Z_{0c}{}^{a} \doteq \frac{-j\zeta_0 A_{c1}{}^{a}}{2\pi} \tag{4.5b}$$

and where, using (2.19c, d),

$$Z_{1in} = \left[\frac{1}{Z_{0c}{}^{s}(1+k^{-1})} + \frac{1}{Z_{0c}{}^{a}(1+k)}\right]^{-1} = Z_{1in}{}' F_1 \tag{4.5c}$$

where

$$Z_{1in}{}' = \frac{2Z_{0c}{}^{s} Z_{0c}{}^{a}}{Z_{0c}{}^{s} + Z_{0c}{}^{a}};$$

$$F_1 = 1 + \frac{[Z_{0c}{}^{s} - Z_{0c}{}^{a}][k - k^{-1}]}{2[Z_{0c}{}^{s}(1+k^{-1}) + Z_{0c}{}^{a}(1+k)]}. \tag{4.5d}$$

It is seen from (4.5) that $I_1(z)$ and Z_{1in} involve only $A_{c1}{}^{s}$, $A_{c1}{}^{a}$, and k. As with the isolated antenna, the former are defined as follows when $\beta_0 h = \pi/2$:

$$A_{c1} \equiv F_{c1}(h) = -\int_{-h}^{h} \cos\beta_0 z' K_{c1}(h, z') dz'; \quad h = \lambda_0/4 \tag{4.6}$$

where

$$K_{c1}(h, z') = K_1(h, z') \pm [K_1(d-h, z') + K_1(d+h, z')]; \quad h = \lambda_0/4. \tag{4.7}$$

Using the integrals given in reference 4, (4.6) may be expressed as follows:

$$A_{c1} = -\{C_a(h, h) \pm [C_a(d-h, h) + C_a(d+h, h)]\}; \quad h = \lambda_0/4. \tag{4.8}$$

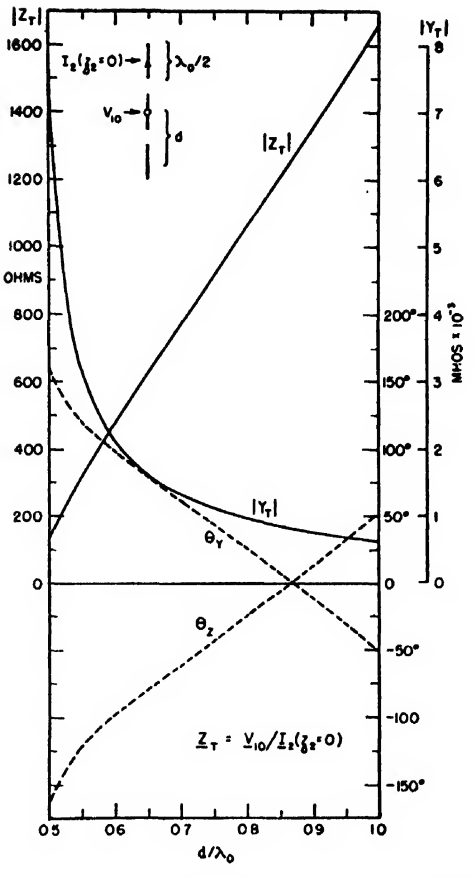


FIG. 10. Transfer impedance and admittance for antenna with two collinear parasites.

The upper sign requires a superscript s, the lower sign a superscript a on A_{c1}. It follows that

$$Z_{0c} = \frac{j\zeta_0}{2\pi}[C_a(h, h) \pm C_a(d-h, h) \pm C_a(d+h, h)]; \quad h = \lambda_0/4 \tag{4.9}$$

where the upper sign is for $Z_{0c}{}^{s}$, the lower sign for $Z_{0c}{}^{a}$. The symmetrical and antisymmetrical impedances of the three element array as given in (4.9) are in Fig. 6. The impedance $Z_{1in}{}'$ defined in (4.5c), which is the principal part of Z_{1in}, is shown in Fig. 9 together with Z_{1in}. Note that Z_{1in} reduces to $Z_{1in}{}'$ when d is sufficiently great so that $k \doteq 1$. It is not possible to determine Z_{1in} until k is available. This requires the evaluation of $Z_{even}{}^{s}$ and $Z_{even}{}^{a}$ in addition to $Z_{0c}{}^{s}$ and $Z_{0c}{}^{a}$.

The first-order symmetrical and antisymmetrical even currents in the outer units are given by (3.19a) with $\beta_0 h = \pi/2$. They are

$$I_{even}{}^{s}(z_2) = \frac{j2\pi V_2{}^{s}}{\zeta_0 A_{even1}{}^{s}} \times [\cos\beta_0 z_2 + M_{even1}{}^{s}(z_2)/\Psi_{even}{}^{s}] \tag{4.10a}$$

$$I_{\text{even}}{}^{a}(z_2)=\frac{j2\pi V_2{}^{a}}{\zeta_0 A_{\text{even1}}{}^{a}}$$

$$\times[\cos\beta_0 z_2+M_{\text{even1}}{}^{a}(z_2)/\Psi_{\text{even}}{}^{a}]. \quad (4.10b)$$

When the outer units are parasitic

$$V_2{}^{s}=V_2{}^{a}=f^{s}V^{s}=f^{a}V^{a}=V_{10}/2F_2;$$

$$F_2\equiv\frac{f^{a}+f^{s}}{2f^{a}f^{s}}=\frac{1}{2}\left[\frac{Z_{0c}{}^{s}}{Z_{\text{even}}{}^{s}}+\frac{Z_{0c}{}^{a}}{Z_{\text{even}}{}^{a}}\right] \quad (4.10c)$$

as in (3.23a) and the total current is given by,

$$I_{2\text{even}}(z_2)=I_{\text{even}}{}^{s}(z_2)-I_{\text{even}}{}^{a}(z_2)$$

$$=\frac{j\pi V_{10}}{\zeta_0 F_2}\left\{\left[\frac{1}{A_{\text{even1}}{}^{s}}-\frac{1}{A_{\text{even1}}{}^{a}}\right]\cos\beta_0 z_2\right.$$

$$\left.+\left[\frac{M_{\text{even1}}{}^{s}(z_2)}{A_{\text{even1}}{}^{s}\Psi_{\text{even}}{}^{s}}-\frac{M_{\text{even1}}{}^{a}(z_2)}{A_{\text{even1}}{}^{a}\Psi_{\text{even}}{}^{a}}\right]\right\};$$

$$\beta_0 h=\pi/2. \quad (4.11)$$

The first-order symmetrical and antisymmetrical impedances of the outer units are

$$Z_{\text{even}}{}^{s}=\frac{-j\zeta_0 A_{\text{even1}}{}^{s}}{2\pi}[1+B_{\text{even1}}{}^{s}/\Psi_{\text{even}}{}^{s}]^{-1};$$

$$Z_{\text{even}}{}^{a}=\frac{-j\zeta_0 A_{\text{even1}}{}^{a}}{2\pi}[1+B_{\text{even1}}{}^{a}/\Psi_{\text{even}}{}^{a}]^{-1}. \quad (4.12a)$$

The transfer impedance when the outer units are parasitic is

$$Z_T=\frac{-j\zeta_0 F_2}{\pi}\left\{\left[\frac{1}{A_{\text{even1}}{}^{s}}-\frac{1}{A_{\text{even1}}{}^{a}}\right]\right.$$

$$\left.+\left[\frac{B_{\text{even1}}{}^{s}}{A_{\text{even1}}{}^{s}\Psi_{\text{even}}{}^{s}}-\frac{B_{\text{even1}}{}^{a}}{A_{\text{even1}}{}^{a}\Psi_{\text{even}}{}^{a}}\right]\right\}^{-1} \quad (4.12b)$$

where

$$B_{\text{even1}}{}^{s}\equiv M_{\text{even1}}{}^{s}(0); \quad B_{\text{even1}}{}^{a}\equiv M_{\text{even1}}{}^{a}(0). \quad (4.12c)$$

For sufficiently great values of $\Psi_{\text{even}}{}^{s}$ and $\Psi_{\text{even}}{}^{a}$ the leading terms are good approximations. They are

$$I_{\text{even}}{}^{s}(z_2)\doteq\frac{V_2{}^{s}}{Z_{\text{even}}{}^{s}}\cos\beta_0 z_2;$$

$$I_{\text{even}}{}^{a}(z_2)\doteq\frac{V_2{}^{a}}{Z_{\text{even}}{}^{a}}\cos\beta_0 z_2;$$

$$I_{2\text{even}}(z_2)=\frac{V_{10}}{Z_T}\cos\beta_0 z_2 \quad (4.13a)$$

where

$$Z_{\text{even}}{}^{s}=\frac{-j\zeta_0 A_{\text{even1}}{}^{s}}{2\pi}; \quad Z_{\text{even}}{}^{a}=\frac{-j\zeta_0 A_{\text{even1}}{}^{a}}{2\pi}, \quad (4.13b)$$

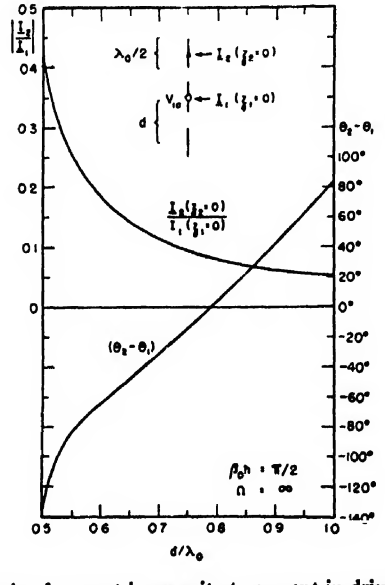


FIG. 11. Ratio of current in parasite to current in driven element.

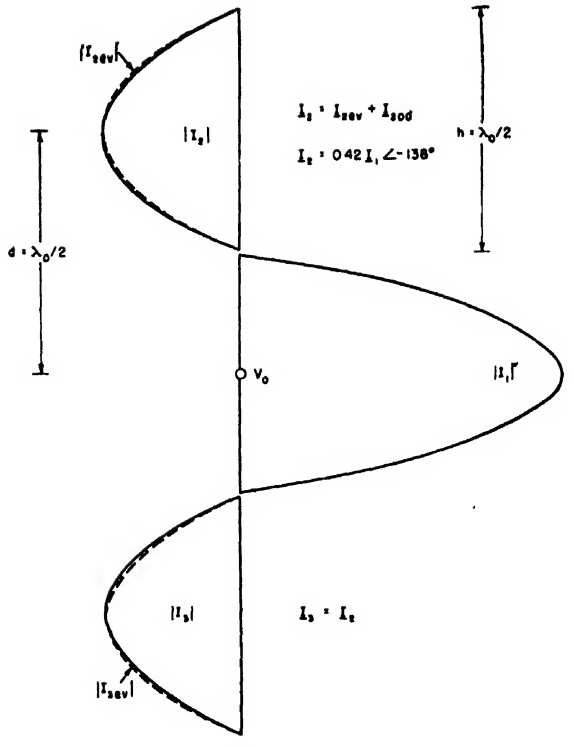


FIG. 12. Distributions of current on collinear array with outer elements parasitic.

$$Z_T = Z_T' F_2; \quad Z_T' \equiv \frac{2Z_{even^s} Z_{even^a}}{Z_{even^a} - Z_{even^s}};$$

$$F_2 \equiv \frac{1}{2}\left[\frac{Z_{0c^s}}{Z_{even^s}} + \frac{Z_{0c^a}}{Z_{even^a}}\right] \doteq 1. \quad (4.13c)$$

The ratio of currents at the centers of the outer parasitic units to the current at the center of the driven unit is given by

$$I_2(z_2=0)/I_1(0) = Z_{11n}/Z_T = \left|\frac{I_2(0)}{I_1(0)}\right| e^{j(\theta_2-\theta_1)}, \quad (4.14)$$

where Z_{11n} is given by (4.5c) and Z_T by (4.13b). It is shown later (Fig. 8) that the factor F_2 differs negligibly from unity as indicated in (4.13c) except when d/λ_0 is at or very near 0.5. As in (4.6), but with a different kernel,

$$A_{even1} = F_{even1}(h) = -\int_{-h}^{h} \cos\beta_0 z_2' K_{1even}(h, z_2') dz_2'$$

$$= -\int_0^h \cos\beta_0 z_2' K_{even}(h, z_2') dz_2' \quad (4.15)$$

where $K_{even}(h, z_2')$ is given by (2.17a) with $z_2 = h = \lambda_0/4$. Thus,

$$K_{even}(h, z_2) = K(h, z_2') \pm \tfrac{1}{2}[K(d-h, z_2') + K(d+h, z_2')] + \tfrac{1}{2}[K(2d-h, z_2') + K(2d+h, z_2')] \quad (4.16)$$

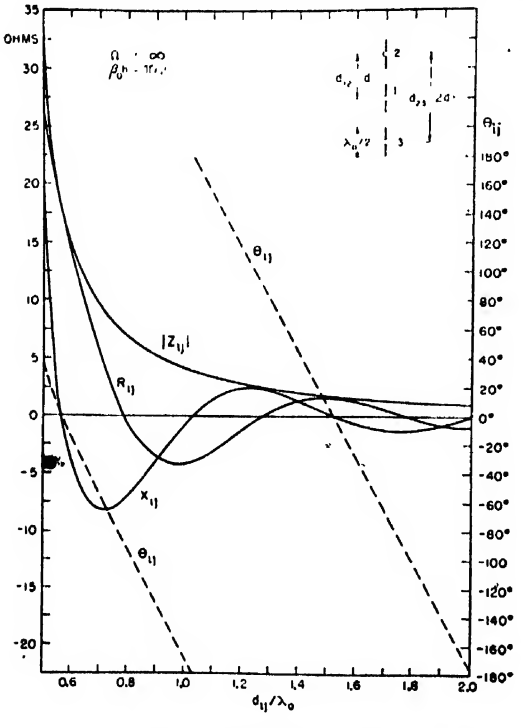


FIG. 13. Mutual impedance for collinear arrays.

where, as usual, $K \equiv K_1 + K_2$. With reference 4 the integrals in (4.13a) with (14.4) may be expressed as follows:

$$A_{even1} = -\{C_a(h, h) \pm \tfrac{1}{2}[C_a(d-h, h) + C_a(d+h, h)] + \tfrac{1}{2}[C_a(2d-h, h) + C_a(2d+h, h)]\}; \quad (h = \lambda_0/4). \quad (4.17)$$

The upper sign in (4.17) goes with a superscript s on A_{even1}, the lower sign with a superscript a. The corresponding impedances are

$$Z_{even} = \frac{j\zeta_0}{2\pi}\{C_a(h, h) \pm \tfrac{1}{2}[C_a(d-h, h) + C_a(d+h, h)] + \tfrac{1}{2}[C_a(2d-h, h) + C_a(2d+h, h)]\}; \quad h = \lambda_0/4, \quad (4.18)$$

where the upper sign is for Z_{even^s}, the lower sign for Z_{even^a}. They are in Fig. 7.

With Z_{even^s} and Z_{even^a} determined, the functions k, f^a and f^s defined in (4.1b) and the function F_2 defined in (4.5d) may be evaluated k and F_2 are in Fig. 8 in polar form. With k available and Z_{11n}' known, the approximate input impedance Z_{11n} (as defined in (4.5c)) may be evaluated. It is shown in Fig. 9 along with Z_{11n}' which it approaches as k approaches unity with increased d. With F_2 available the transfer impedance Z_T may be determined using its principal factor Z_T'. Z_T and Z_T' are in Fig. 10. Note that $Z_T \doteq Z_T'$ except when $d/\lambda_0 \doteq 0.5$, and even here the departure is small. The ratio of currents (4.14) is shown in Fig. 11 as a function of d/λ_0 and the distribution in Fig. 12 for $d = \lambda_0/2$. Note that when $d = \lambda_0/2$ the currents in the parasites are nearly opposite in phase to the current in the central unit.

The general equations for the three-element array in which the outer units are always identical are

$$V_{10} = I_{10}Z_{s1} + I_{20}Z_{12} + I_{30}Z_{13} = I_{10}Z_{s1} + 2I_{20}Z_{12} \quad (4.19a)$$

$$V_{20} = I_{10}Z_{21} + I_{20}Z_{s2} + I_{30}Z_{23} = I_{10}Z_{12} + I_{20}(Z_{s2} + Z_{23}). \quad (4.19b)$$

When the units are symmetrically driven,

$$V^s = I_{10}(Z_{s1} + 2Z_{12}), \quad (4.20a)$$

$$V_2^s = I_{10}(Z_{s2} + Z_{12} + Z_{23}). \quad (4.20b)$$

When the units are antisymmetrically driven,

$$V^a = I_{10}(Z_{s1} - 2Z_{12}), \quad (4.21a)$$

$$V_2^a = -I_{10}(Z_{s2} - Z_{12} + Z_{23}). \quad (4.21b)$$

Comparison with (9) and (18) shows that when $\beta_0 h = \pi/2$ these approximate expressions for the symmetrical and antisymmetrical impedances are consistent with the following definitions:

$$Z_{s1} \doteq Z_{s2} \doteq (j\zeta_0/2\pi) C_a(h, h), \tag{4.22a}$$

$$Z_{12} = Z_{13} \doteq \frac{j\zeta_0}{2\pi}\frac{1}{2}[C_a(d-h, h) + C_a(d+h, h)], \tag{4.22b}$$

$$Z_{23} \doteq \frac{j\zeta_0}{2\pi}\frac{1}{2}[C_a(2d-h, h) + C_a(2d+h, h)]. \tag{4.22c}$$

The functions Z_{12} and Z_{23} may be obtained from Fig. 13. It may be concluded that in general, when $\beta_0 h = \pi/2$, the zeroth-order self-impedance of each of N collinear antennas is

$$Z_{s1} = \cdots Z_{sN} \doteq \frac{j\zeta_0}{2\pi} C_a(h, h) = 73.13 + j42.5 \text{ ohms}. \tag{4.23a}$$

Similarly the zeroth-order mutual impedances between two collinear antennas numbered m and n and separated a distance $(n-m)d$ between centers, is

$$Z_{mn} \doteq \frac{j\zeta_0}{2\pi}\frac{1}{2}\{C_a[(n-m)d-h, h] + C_a[(n-m)d+h, h]\}. \tag{4.23b}$$

For collinear elements that are *not* near $\beta_0 h = \pi/2$ in electrical half-length the self- and mutual impedances cannot be expressed in this simple form. Only when $\beta_0 h = \pi/2$ and the identical antennas are sufficiently thin so that Ψ_{even^s} and Ψ_{even^a} are large compared with unity are the simple relations (4.22a, b, c) and (4.23a, b) moderately good approximations. In general, if the antennas are sufficiently thin so that (4.23a) is a good approximaion, it may be assumed that (4.23b) is an even better approximation. Tables of the several zeroth-order impedances involved in the three-element collinear array are in Table I.

The odd currents in the parasitic elements are evaluated in Section 3 when $\beta_0 h = \pi/2$. The approximate expression given in (3.40) is

$$I_{2odd}(z_2) = \frac{V_{10}}{\Psi_{odd}} K \sin 2\beta_0 z_2 \tag{4.24a}$$

where

$$K \equiv \frac{\pi}{4}\left[\frac{D^s(d)}{Z_{0c^s}(1+k^{-1})} + \frac{D^a(d)}{Z_{0c^a}(1+k)}\right] \tag{4.24b}$$

where $D^s(d)$ and $D^a(d)$ are given by (3.38b) respectively with the upper and lower sign, and where k is defined in (4.22). It is now readily verified that with $\beta_0 h = \pi/2$ the odd part of the current in the parasites

TABLE I. Zeroth-order impedances for three-element collinear array.

$\beta_0 d/\pi$	Z_{0c^s}	Z_{0c^a}	Z_{even^s}	Z_{even^a}
1.0	126 $+j82.8$	20.3 $+j$ 2.22	95.3 $+j62.0$	42.6 $+j21.6$
1.1	114 $+j47.2$	32.6 $+j37.8$	90.3 $+j46.4$	49.5 $+j41.7$
1.2	102 $+j34.5$	43.8 $+j50.5$	86.4 $+j41.0$	56.7 $+j49.0$
1.3	92.4 $+j28.4$	53.8 $+j56.6$	83.2 $+j38.0$	63.9 $+j52.0$
1.4	84.0 $+j26.2$	62.2 $+j58.8$	80.0 $+j35.8$	69.2 $+j52.0$
1.5	77.1 $+j26.5$	69.1 $+j58.6$	76.8 $+j34.8$	72.8 $+j50.8$
1.6	72.0 $+j28.8$	74.3 $+j56.2$	73.8 $+j35.0$	75.0 $+j48.6$
1.7	68.2 $+j31.4$	78.0 $+j53.6$	71.2 $+j35.7$	76.1 $+j46.8$
1.8	65.9 $+j34.7$	80.3 $+j50.4$	69.3 $+j37.4$	76.5 $+j45.2$
1.9	64.9 $+j38.0$	81.4 $+j47.0$	68.2 $+j39.6$	76.4 $+j44.0$
2.0	64.8 $+j41.0$	81.4 $+j44.0$	68.0 $+j41.8$	76.3 $+j43.2$

$\beta_0 d/\pi$	Z_{12}	Z_{23}	Z_{lin}	Z_T
1.0	26.4 $+j20.2$	-4.16 $-j$ .725	53.2 $+j23.6$	-131 $-j$ 41.8
1.1	20.2 $+j$ 2.35	-3.00 $+j1.56$	63.6 $+j46.0$	-157 $-j$ 278
1.2	14.6 $-j$ 4.00	-1.27 $+j2.50$	71.2 $+j47.4$	$-$ 72.0 $-j$ 462
1.3	9.66 $-j$ 7.05	0.437 $+j2.46$	74.0 $+j45.6$	103.4 $-j$ 618
1.4	5.44 $-j$ 8.14	1.48 $+j1.36$	74.8 $+j44.1$	356 $-j$ 688
1.5	2.00 $-j$ 8.01	1.73 $+j$ .289	74.6 $+j42.4$	638 $-j$ 632
1.6	-0.588 $-j$ 6.84	1.32 $-j$ .747	74.0 $+j42.0$	947 $-j$ 474
1.7	-2.45 $-j$ 5.56	0.538 $-j1.24$	73.2 $+j41.6$	1190 $-j$ 164
1.8	-3.60 $-j$ 3.92	$-$ .184 $-j1.16$	72.5 $+j42.2$	1310 $+j$ 266
1.9	-4.12 $-j$ 2.24	$-$ .805 $-j$ .704	72.7 $+j42.1$	1280 $+j$ 782
2.0	-4.16 $-j$ .725	$-$ .958 $+j$ .017	73.2 $+j42.8$	1020 $+j1290$

$Z_{s1} = Z_{s2} = Z_{s3} = 73.1 + j42.5$

is quite small compared with the even parts. Thus, when $\beta_0 d = \pi$, $h/d = \frac{1}{2}$,

$$D^s(d) = \frac{1}{\pi^2}\left(-\frac{4}{3} + \frac{8}{15}\right) = -0.08, \tag{4.25a}$$

$$D^a(d) = \frac{1}{\pi^2}\left(\frac{4}{3} + \frac{8}{15}\right) 0.19. \tag{4.25b}$$

Also,

$$Z_{0c^s}(1+k^{-1}) = 196.9\underline{/28°.3}; \qquad Z_{0c^a}(1+k) = 82.7\underline{/22.2°}. \tag{4.26}$$

With these values,

$$K \doteq 1.56\times 10^{-3}\underline{/-21°} \tag{4.27}$$

so that with $\Omega = 12.5$, $\Psi_{odd} = 10$,

$$I_{odd}(z_2) = 1.56\times 10^{-4} V_{10} e^{-j21°} \sin 2\beta_0 z_2. \tag{4.28}$$

The corresponding value for the even current with $\beta_0 d = \pi$ is obtained from (4.13a) with $Z_T = 137.7\underline{/-162°}$.

$$I_{even}(z_2) = \frac{V_{10}}{Z_T}\cos\beta_0 z_2 = 72.8\times 10^{-4} V_{10} e^{j162°}\cos\beta_0 z_2. \tag{4.29}$$

Combining (4.28) with (4.29) gives

$$I_2(z_2) = 72.8\times 10^{-4} V_{10} e^{j162°} \times [\cos\beta_0 z_2 - e^{-j3°} 0.022 \sin 2\beta_0 z_2]. \tag{4.30}$$

It is seen from (4.30) that the amplitude of the odd current is only 2 percent of the amplitude of the even current for $\Omega = 12.5$. For thicker antennas the odd currents are relatively greater. The current in (4.30) is shown in Fig. 12 together with the current in the central unit.

5. THE CENTER-DRIVEN COLLINEAR ARRAY WITH THE PHASE-REVERSING STUBS

It is shown in Section 4 that the currents in the two outer units of a three-element collinear array are nearly *opposite* in direction and equal in magnitude to the current in the central unit if the distance between adjacent ends of the elements is very small. The electromagnetic field of such a distribution of current is characterized by large maxima both along the plane bisecting the antenna and along symmetrical cones in the hemispheres. Between these maxima are deep minima. Whereas such a pattern is useful for certain applications, it is undesirable when a field is required that has a large amplitude only in the central plane. It is known that such a field is maintained by a collinear array if the currents are all about equal and in phase. Such a distribution can be maintained in the antennas if each unit is center-driven but, in practice, this involves long transmission lines in non-neutral planes. A better way is to drive only the central unit and to excite the outer units by coupling networks that connect the adjacent ends of the several units. The function of such networks is to *reverse the phase* of the currents in the outer antennas.

A simple phase-reversing network is a high-impedance, transmission-line stub. In Fig. 14a a three-element array is shown with two *phase-reversing stubs* constructed of low loss two-wire line. *If properly adjusted* in length, the transmission-line impedance of each stub to *equal and opposite currents* is very great.

The approximate determination of the current and impedance of the array in Fig. 14a may be carried out as for the array with parasitic elements. The analysis is resolved into two parts the superposition of which yields the desired result. These are (1) the symmetrical part with all three units individually center-driven by slice generators with voltages that maintain equal currents in phase at the centers of the antennas, and (2), the antisymmetrical part in which the driving voltages are so chosen that the currents at the centers of the outer antennas are reversed. These two parts are considered in turn and combined later.

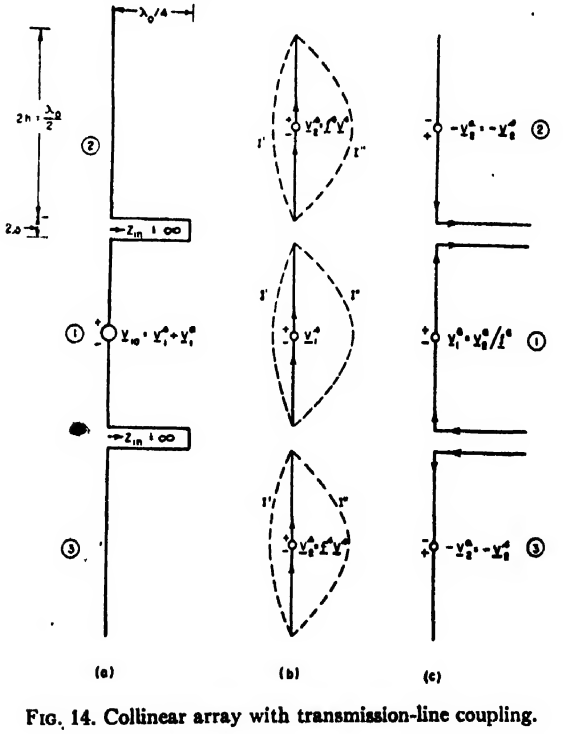


FIG. 14. Collinear array with transmission-line coupling.

The Symmetrical Problem

Let the voltages required to maintain equal currents at the centers of all three units, $I_1^s(0)=I_2^s(z_2=0)=I_3^s(z_3=0)$, be represented as in (3.20a)

$$V_{10}=V_1^s=V^s;\quad V_{30}=V_{20}=V_2^s=f^sV^s. \tag{5.1}$$

It is shown readily that f^s in (5.1) is effectively the same function as in (3.20b) even though the array differs from that involved in Section 3 by the presence of the stubs. Since the currents that enter the stubs from the attached antennas are essentially equal and opposite, and since the stubs are adjusted to offer an extremely high impedance to equal and opposite currents, the currents actually existing at the ends of the antennas where they are connected to the stubs are very small. This does not mean that transmission-line currents all along the stubs are as small as the input currents. Actually, the distribution of transmission-line currents on the high-impedance stubs is approximately sinusoidal with a small amplitude at the input end and a large amplitube in the terminating bridge. However, with closely spaced good conductors the ohmic losses in the stubs and the radiation losses from them are insignificant compared with radiation from the array as a whole, and they contribute nothing to the far-zone field. Indeed, there can be no significant change in either the input impedance of the array or the distribution of current in the three antennas if the stubs are removed as in Fig. 14b, leaving the circuit already analyzed in Section 4 where it constitutes the symmetrical part of the collinear array with parasitic outer units. The symmetrical current $I_1^s(z)$ in the central unit is given by (4.2a) or (4.5a); the symmetrical impedance Z_{0c}^s of the central unit is given by (4.9a) and graphically in Fig. 6. The currents in the outer units are well approximated by (4.10a) or (4.13a) since the odd components of current are negligible in determining the far-zone field of the array. The approximate currents are

$$I_1^s(z)=(V^s/Z_{0c}^s)\cos\beta_0 z \tag{5.2a}$$

$$I_3^s(z_3)=I_2^s(z_2)=(V_2^s/Z_{even}^s)\cos\beta_0 z_2$$
$$=(V^sf^s/Z_{even}^s)\cos\beta_0 z_2=I_1^s(z) \tag{5.2b}$$

where Z_{0c}^s is in (4.9a), Z_{even}^s is in (4.18a), and $f^s=Z_{even}^s/Z_{0c}^s$.

The Antisymmetrical Problem

Let the driving voltages required to maintain the currents $I_3^a(z_3=0)=I_2^a(z_2=0)=-I_1^a(0)$ in the array in Fig. 14a be represented as in (3.21a):

$$V_{10}=V_1^a=V^a;\quad V_{30}=V_{20}=-V_2^a=-f^aV^a. \tag{5.3}$$

The function f^a in (3) is *not* the same as in (3.21b). Since the currents in the two parallel conductors of the phase-reversing stubs are nearly equal and codirectional, these two conductors are always at the same potential at opposite points. Therefore, the terminating bridge may be removed (or replaced by an equivalent open end) without significant effect. Thus, the solution of the antisymmetrical problem of Fig. 14a reduces to the analysis of Fig. 14c. Unfortunately, this presents an intricate problem in a quantitative sense except for conductors of very small radius. Nevertheless, much valuable information on the general nature of the distribution of current and the operation of the array, and even a fair estimate of the impedance can be acquired from a rough quantitative investigation.

It has been shown[6] for parallel antennas that the current in an antenna that is coupled to another is determined in *major outline* by the dimensions of the antenna itself even if the coupling is quite close. Therefore, the first step in the study of Fig. 14c is directed toward the behavior *when isolated* of each of the three antennas with its attached half of the now open stub. It is reasonable to assume that the effects of the right-angled bends in the conductors can, at most, lead to a small quantitative modification, but not to essential changes in the distributions of current and charge. This is indicated from the study of the V-antenna.[6] Accordingly, the distributions of current in antennas 1 and 2 of Fig. 14c must be approximately the distributions in the same conductors when these are straightened as in Fig. 15. This makes the central unit equivalent to a center-driven antenna of electrical half-length $\beta_0 h=\pi$. The distribution of current on such an antenna is reproduced in Fig. 15f, and transferred to the bent antenna in Fig. 15e as an approximation. The impedance is simply

$$Z_{0c}^a=Z_0(\beta_0 h=\pi)\equiv Z_\pi \tag{5.4}$$

The bent antenna 2 in Fig. 15a may be approximated by the asymmetrically driven straight antenna in Fig. 15b. The analysis of this[7] has shown that in solving for the current in either half, this half may be assumed to be erected on an infinite, perfectly conducting plane and driven by a slice generator with appropriate voltage. For the upper half with an electrical length $\beta_0 h=\pi/2$ and an impedance $\frac{1}{2}Z_0(\beta_0 h=\pi/2)\equiv(Z_{\pi/2})/2$ the driving voltage is

$$V_{\pi/2}\equiv -V_2^aZ_{\pi/2}/(Z_\pi+Z_{\pi/2}); \tag{5.5a}$$

for the lower half (which has an electrical length $\beta_0 h=\pi$ and an impedance $\frac{1}{2}Z_0(\beta_0 h=\pi)\equiv(Z_\pi)/2$) the driving voltage is

$$V_\pi=-V_2^aZ_\pi/(Z_{\pi/2}+Z_\pi). \tag{5.5b}$$

Since the two antennas are in series, the resultant driving voltage is $-V^a$ as assumed in (5.3). The impedance is

$$Z_2^a=\tfrac{1}{2}(Z_\pi+Z_{\pi/2}). \tag{5.6}$$

The distributions of current on the two dissimilar parts of the outer antennas are shown in Fig. 15c, d with phases referred to $-V_2^a$. It is assumed that these distributions for isolated straight asymmetrical antennas approximate the distributions on the bent antenna in Fig. 15a.

With approximate distributions of current in the three antennas of Fig. 14c determined under the assumption that they are isolated, it remains to evaluate the effect of their interaction. An estimate of this may be obtained as follows. The most closely coupled parts of antennas 1 and 2 are the parallel ends at right angles to the axis of the array. These parts carry co-directional currents. It is known that the effect upon a given isolated antenna of a closely spaced parallel antenna with equal co-directional current is to increase the expansion parameter somewhat. That is, $\Psi_{\pi 1}$ for two symmetrically driven, parallel antennas is greater than Ψ_{K1} for the isolated antenna if b/λ_0 is small. For close

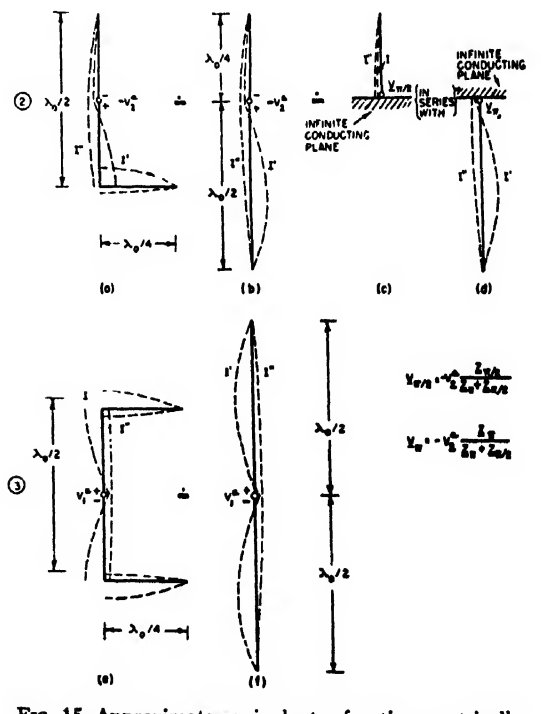


FIG. 15. Approximate equivalents of antisymmetrically driven antennas in three-element collinear array.

[5] C. T. Tai, Proc. I.R.E. 36, 487–500 (1948).
[6] R. King, "Theory of V-Antennas," Cruft Laboratory Technical Report No. 96 (February, 1950).
[7] R. King, "Asymmetrically driven antennas and the sleeve dipole," Cruft Laboratory Technical Report No. 93, (December, 1949). Proc. I.R.E. 38, 1154–1164 (1950).

spacing this increase is of the order of magnitude of 3 to 6. For thick antennas with $\Omega=2\ln(2h/a)$ not very large, the percentage change in Ψ_{K1} is great; for very thin wires it is quite small. The collinear parts of antennas 1 and 2 carry currents that are practically equal and opposite. It has been shown in Section 2 that a coupled antenna carrying an equal and opposite currents reduces the expansion parameter approximately in the same proportion as it is increased when the currents are co-directional. Thus, antenna 1 in the proximity of antennas 2 and 3 in Fig. 14c behaves roughly like a straight isolated antenna of the same length of conductor which has outer halves that are effectively of smaller radius, inner halves that are effectively of greater radius than when isolated. The same reasoning applies to the outer antennas.

To summarize, it is evident that a good estimate of the distribution of current and of the driving-point impedance of the three-element collinear array in Fig. 14a may be obtained by superimposing the solution of the symmetrically driven array in Fig. 14b upon the solution of the antisymmetrically driven array in Fig. 14c. In this latter, the currents in the central unit are like that in Fig. 15f. The resultant current is the algebraic sum of these symmetrical and antisymmetrical currents. In addition, there is the transmission-line distribution of equal and opposite currents on the stubs with large amplitude in the connecting bridges. Evidently, the method applied to the three-element array may be extended to any number of units. Note that the approximations involved improve as the value of $\Omega=2\ln(2h/a)$ increases.

Since the amplitude of the current in an antenna of electrical half-length $\beta_0 h=\pi$ is very much smaller than in an antenna for which $\beta_0 h=\pi/2$ when both are driven by the same voltage (for $\Omega=20$ the ratio of maximum currents is roughly 1 to 12), and since the antisymmetrically driven array in Fig. 15c corresponds roughly to elements with $\beta_0 h=\pi$, whereas the symmetrically driven array in Fig. 15b corresponds roughly to elements with $\beta_0 h=\pi/2$, it follows that even for moderately large ratios h/a, the predominating currents are the symmetrical currents of Fig. 14b as given in (5.1). *It is primarily these currents that determine the far-zone electromagnetic field.* Note that they are co-directional in the several units. That is, the phase-reversing stubs have, indeed, effectively reversed the *significantly radiating currents* in the outer units. Note that additional currents are maintained both on the stubs and on the antennas, but their principal role is to maintain the desired co-directional currents on the antennas while contributing negligibly to the radiation field either because they are small (viz., additional currents on antennas and co-directional currents on stubs) or, if large, because they are equal, opposite, and close together (as with the transmission-line currents on the stubs).

The driving-point admittance of the central unit and, hence, of the entire array is formally like (2.19b), viz.,

$$Y_{in}=\frac{I_1(0)}{V_{10}}=\frac{I_1^e(0)+I_1^a(0)}{V^e+V^a}$$

$$=\frac{1}{Z_{0c}^e(1+k^{-1})}+\frac{1}{Z_{0c}^a(1+k)}, \quad (5.7)$$

where Z_{0c}^e is given by (4.9a) and is plotted in Fig. 6, $Z_{0c}^a=Z_\pi$, and k is defined by

$$k\equiv\frac{f^a}{f^e};\quad f^a=\frac{Z_2^a}{Z_{0c}^a}=\frac{Z_\pi+Z_{\pi/2}}{2Z_\pi};\quad f^e=\frac{Z_2^e}{Z_{0c}^e}=\frac{Z_{\text{even}}^e}{Z_{0c}^e} \quad (5.8)$$

where Z_{even}^e is given by (4.18a) and is plotted in Fig. 7. If the separation b of the conductors of the transmission-line stubs is very small, $d/\lambda_0\doteq 0.5$. For this value $Z_{0c}^e\doteq 126+j83$, $Z_{\text{even}}^e\doteq 95+j62$; $f^e\doteq 0.75$. For sufficiently thin antennas, $|Z_\pi|\gg|Z_{\pi/2}|$ so that $f^a\doteq 0.5$. Hence $k\doteq\frac{2}{3}$. With this value,

$$Z_{in}=\frac{1}{Y_{in}}=\frac{Z_{0c}^e Z_\pi(1+\frac{2}{3})(1+\frac{3}{2})}{Z_{0c}^e(1+\frac{3}{2})+Z_\pi(1+\frac{2}{3})}. \quad (5.9)$$

Since for thin antennas, $|Z_\pi|\gg|Z_{0c}^e|$, it follows that

$$Z_{in}\doteq Z_{0c}^e(1+\tfrac{3}{2})=2.5Z_{0c}^e=2.5(126+j83)$$

$$=315+j207. \quad (5.10)$$

This value corresponds to $73.1+j42.5$ for a single unit. More accurate numerical results may be obtained for specific values of Z_π and $Z_{\pi/2}$ if desired. Other values of d/λ_0 may be used provided the spacing b of the stub wires satisfies $\beta_0 b\ll 1$.

Although the investigation of the three-element center-driven collinear array with phase-reversing stubs is not rigorous, it does provide an excellent picture of the several components of current and their distributions. Moreover, it justifies the conventional practice of assuming currents in the elements of the array co-directional, and sinusoidally distributed with equal amplitude in determining radiation-field patterns with $\beta_0 h=\pi/2$. That is, the dominant, symmetrical distribution in Fig. 14b approximates the distribution

$$I_s\doteq I_0\cos\beta_0 z_i;\quad i=1,2,3 \quad (5.11)$$

where z_i is measured from the center of antenna i.

The advantages of the collinear array with phase-reversing stubs over the same antennas with each element center-driven from a transmission line are clear. In the first place, it is much simpler from the constructional point of view. In the second place, it provides a far-zone field that is determined entirely by currents in the antennas and not by currents in the feeding lines. While it is true that the stub sections of line have unbalanced currents, it has been shown that these are

small compared with the principal currents on the antennas. Moreover, the stubs are of fixed and relatively short length so that contributions from small unbalanced currents in them are both negligible and determinate. Since individual transmission lines usually are long, and unbalanced currents in them may be large if the over-all electrical length of the line and attached metal at the generator-end happen to be self-resonant, their contribution to the far-zone field is neither determinate nor is it usually small.

Beam Oscillations in an F-M Cyclotron*

T. TEICHMANN
Princeton University, Princeton, New Jersey
(Received May 8, 1950)

This paper considers the radial and vertical oscillations of a synchro-cyclotron beam. It is found that various oscillations of substantially constant or increasing amplitude may occur in certain regions of the field, in which there are angular or other asymmetries, though only one possible oscillation (that at $n\simeq\frac{1}{4}$) can occur in a field that is cylindrically and vertically symmetric. The conditions for the various oscillations, and their behavior, are discussed.

I. INTRODUCTION

THE aim of this paper is to give a general discussion of various oscillations of a cyclotron beam which are induced (or permitted) by the particular shape or angular inhomogeneity of the magnetic field.[1]

The calculation is given only for small oscillations (i.e., oscillations of amplitude small compared to the radius of the beam) and depends on the adiabatic approximation, i.e., on the smallness of the fractional gain of energy of the beam over the period of several oscillations. The discussion is substantially non-relativistic, but can easily be extended to small relativistic energies.

II. NOTATION

One thus considers the motion of a particle of charge e, mass m, and energy $E=mv^2/2$, in a temporally constant magnetic field with cylindrical components (H_r, H_θ, H_z). Because of the particular design of cyclotrons, it is possible to find a median plane (usually parallel to and midway between the pole faces of the magnet) and a perpendicular axis, such that the field may be represented in good approximation by the cylindrical-components $(0, 0, H_n)(H_n=\text{constant})$. This plane and axis are then taken as the basis for a cylindrical coordinate system to describe the position of the particles (r, θ, z) and the spatial variation of the field. If the field were exactly $(0, 0, H_n)$ and the particle were set moving in the median plane, the subsequent motion would be of constant angular frequency

$$w_n=eH_n/mc \tag{2.1}$$

in a circle of radius

$$r_n=v/w_n. \tag{2.2}$$

In practice, because of the field variation, and possible angular inhomogeneities, the motion will be almost periodic (instead of strictly periodic) and may be regarded as uniform circular motion with superposed vertical and radial oscillations about the instantaneous equilibrium orbit (to be denoted by "i.o." in the following). The accelerating field of the dees will, of course, gradually increase the velocity v, and hence the i.o., but over a large portion of the particles path in the synchro-cyclotron this increment is relatively small over several periods of the oscillations considered, and the i.o. radius r_n may be regarded as constant in the region of any particular oscillation. If the oscillations about the equilibrium have a very large period, this (adiabatic) approximation loses its validity (see below).

It is convenient to refer all quantities to the i.o. and to introduce the new (non-dimensional) coordinates,

$$x=(r-r_n)/r_n \tag{2.3}$$

$$y=z/r_n, \tag{2.4}$$

and the velocity v_n of motion in the i.o. with angular velocity w_n,

$$v_n=-r_nw_n. \tag{2.5}$$

The magnetic field $\mathbf{H}=(H_r, H_\theta, H_z)$ is then most conveniently written in the form

$$\mathbf{H}=H_n(R, Q, Z) \tag{2.6}$$

(where $R\simeq0$, $Q\simeq0$, $Z\simeq1$). The (non-dimensional) components R, Q, Z, may be expanded in double Taylor series in x, y about any equilibrium radius r_n, and in Fourier series in θ. Carrying the Taylor series up to second-degree terms (i.e., using a parabolic approximation)

$$R=R_0+R_xx+R_yy+\tfrac{1}{2}R_{xx}x^2+R_{xy}xy+\tfrac{1}{2}R_{yy}y^2 \tag{2.7}$$

* This work was assisted in part by the joint program of the ONR and AEC.

[1] The discussion is thus an extension to other regions of the field of considerations given by Hamilton and Lipkin (unpublished) for $n\simeq\frac{1}{4}$.

and Q, Z may be expressed in the same way. Further,

$$R_y=\sum_{-\infty}^{+\infty} R_y{}^k e^{ik\theta}, \text{ etc.} \tag{2.8}$$

Note that H_n is a parameter which is regarded as constant over the domain of any oscillation, but which varies from i.o. to i.o. (as r_n increases). The normalization is so chosen that

$$Z_0=1+\sum_{k\neq 0} Z_0{}^k e^{ik\theta}. \tag{2.9}$$

Because of the electromagnetic equations curl$\mathbf{H}=0$, and div$\mathbf{H}=0$, and possible symmetries, the coefficients defined above satisfy certain linear equations which are listed in the Appendix, and used in the remainder of this paper. The most important of these relations are the following:

$$\int_{-\pi}^{\pi} Q d\theta=0, \text{ implying } Q_\alpha{}^0=0, \text{ for all } \alpha=x, y \text{ etc.} \tag{2.10}$$

$$R_y=Z_x \tag{2.11}$$

$$R_{xy}=Z_{xx} \tag{2.12}$$

$$R_{yy}=Z_{xy}. \tag{2.13}$$

If the vertical component of the field is symmetric about the median plane, one has the additional relations

$$Z_y=Z_{xy}=R_{yy}=0. \tag{2.14}$$

The usual definition

$$n=-d\log H_z/d\log r \tag{2.15}$$

is also slightly modified here to apply only to the angularly symmetric part of H_z, i.e.,

$$Z_x{}^0=-nZ_0{}^0=-n \tag{2.16}$$

and hence,

$$Z_{xx}{}^0=(n^2+n-n'), \quad n'=dn/dx. \tag{2.17}$$

III. SUMMARY OF RESULTS

It is found that approximately periodic oscillations about the i.o. can occur only in the neighborhood of certain particular values of n, and that the possibility of developing an appreciable amplitude then depends not only on the shape of the field, but also on the initial phase ψ_n which is in general determined both by the field shape and by the initial displacement of the particle from the i.o., and its velocity. When they do occur, these oscillations have the form†

$$A(\theta)\cos(c\theta+\phi) \tag{3.1}$$

where $\theta\simeq w_n t$ is the path length, and $A(\theta)$ is a slowly varying function of θ. Only the cases in which $A(\theta)$ can remain substantially constant, or increase, are of practical interest.

(i) There are only two oscillations which can occur if the field is cylindrically symmetric, the most important of these being a coupled radial-vertical oscillation which occurs near $n\simeq\frac{1}{3}$, and is fully discussed in reference 1. If X and Y are the amplitudes of the oscillations of x and y (i.e., the ratios of the amplitudes of the radial and vertical oscillations to the i.o. radius r_n), then

$$4X^2+Y^2=\text{constant}, \tag{3.2}$$

so that the oscillations are substantially limited by their initial values, though a radial deviation will turn into a vertical one in the course of time and conversely.

(ii) The other possible oscillations (except at $n\simeq\frac{1}{3}, \frac{1}{5}$) depend on the presence of fundamental or second harmonic terms in the Fourier expansions of R, Q and Z. Proceeding systematically in the direction of increasing n, there is first a purely vertical oscillation at $n\simeq\frac{1}{2}$. For small θ the amplitude has the form

$$Y=Y_0/(1-Y_0|C|\cos u\cdot\theta), \tag{3.3}$$

Y_0 being the initial amplitude (deviation from i.o.), and C a certain linear combination of first Fourier coefficients [see Eq. (5.4)]. Since Y_0, $|C|$ are positive by definition, unpleasant oscillations are only likely if $\cos u>0$.

(iii) At $n\simeq\frac{1}{4}$ a purely vertical oscillation may again occur, though of rather different form, to wit,

$$Y=Y_0\exp(|D|\cos u\cdot\theta) \tag{3.4}$$

for small θ. D is substantially the fundamental Fourier coefficient of Z_z, i.e., of dH_z/dr. Here again nasty oscillations may occur if $\cos u>0$.

(iv) At $n\simeq 9/25$ coupled radial-vertical oscillations may occur both if certain fundamental Fourier coefficients and/or if certain second harmonics are present in the field, though if both types of terms are present together the equations are rather too complicated to be solved in an informative manner. In each separate case the amplitudes satisfy relations of the type

$$|M|X^2\pm|L|Y^2=\text{constant}, \tag{3.5}$$

the quantities L and M being linear combinations of the Fourier coefficients involved. If the plus sign obtains, the oscillations are unlikely to become obtrusive, since they are substantially limited by their initial values, though coupling from radial to vertical oscillation, or the converse, may give rise to trouble, especially if L and M differ widely in magnitude. If the minus sign obtains, on the other hand, the oscillations may grow very large, since it then turns out that X and Y may be represented approximately by hyperbolic functions of θ [see Eq. (5.24)ff].

(v) When $n\simeq 4/9$ there is again the possibility of a purely vertical oscillation, of similar form to that at $n\simeq\frac{1}{3}$, the only difference being that the coefficient C

† The frequencies of the oscillations discussed below are respectively, $w_n(1-n)^{\frac{1}{2}}$ and $w_n(n)^{\frac{1}{2}}$ for the radial and vertical oscillations.

now depends on the second harmonics in the field components.

(vi) At $n\simeq\frac{1}{2}$ there is the possibility of a coupled oscillation if the field is not vertically symmetric about the median plane, quite apart from any angular inhomogeneity. The amplitudes satisfy the relation

$$X^2+Y^2=\text{constant}, \tag{3.6}$$

so that the oscillations are not likely to be troublesome. For small θ, X and Y are themselves approximately periodic in θ with period

$$2\pi/B\cos u. \tag{3.7}$$

Proceeding to larger values of n, one finds that oscillations may occur for those values of $1-n$ for which oscillations occur at n (as above), with the difference that the roles of radial and vertical oscillations are interchanged. It is not of interest to discuss these further, except to remark that the oscillation at $n\simeq\frac{4}{5}$, which is the analog of that at $n\simeq\frac{1}{5}$, here depends on the existence of an asymmetry about the median plane [by virtue of (2.13)], and hence may be avoided in practice.

The upshot of these considerations is that by careful design of the vertical and cylindrical symmetry of the field it is possible to get rid of all the oscillations of the beam, except that at $n\simeq\frac{1}{5}$, which is by its nature generally of limited amplitude.

IV. EQUATION OF MOTION OF THE BEAM (PARTICLES)‡

The motion of a particle of charge e and mass m, in a magnetic field (H_r, H_θ, H_z) is determined by the equations

$$\dot{v}_r=\ddot{r}-r\dot{\theta}^2=(e/mc)(v_\theta H_z-v_z H_\theta) \tag{4.1}$$

$$\dot{v}_\theta=r\ddot{\theta}+2\dot{r}\dot{\theta}=(e/mc)(v_z H_r-v_r H_z)+\text{accelerating force} \tag{4.2}$$

$$\dot{v}_z=\ddot{z}=(e/mc)(v_r H_\theta-v_\theta H_r). \tag{4.3}$$

Instead of Eq. (4.2) it is more convenient to use

$$v^2=v_r^2+v_\theta^2+v_z^2 \tag{4.4}$$

which is essentially the energy integral. The effect of the accelerating force is to change v as the motion progresses, and the adiabatic approximation is permissible provided that this change is relatively small during the time of one radial or vertical oscillation, whichever is the larger; a small relative change per revolution of the particle is not enough. Consequently the method below should not be taken too seriously for n near 0 or 1, unless the fractional increase in energy is exceedingly small at these values.

In terms of the quantities w_n, r_n, and the new variables x, y, the equations of motion become

$$\ddot{x}=w_n(v_\theta Z/r-\dot{y}Q)+v_\theta^2/rr_n \tag{4.5}$$

$$\ddot{y}=w_n(\dot{x}Q-v_\theta R/r) \tag{4.6}$$

and (4.4) (since $v_r=r_n\dot{x}$, $v_\theta=r\dot{\theta}$, $v_z=r_n\dot{y}$). It is now convenient to introduce the new variable $\vartheta=w_n t$. In view of the smallness of the oscillations and the adiabatic principle, $\vartheta\simeq\theta$. Actually they are related by $\vartheta/\theta=w_n/\theta$, and $w_n/\theta\simeq 1$. In the following it will be assumed that $\vartheta=\theta$. x' then denotes $dx/d\theta$, etc. It is also convenient to write

$$v^2=v_n^2(1+\epsilon) \tag{4.7}$$

(i.e., the total energy $mv^2/2=$energy of rotation in an idealized i.o. $mv_n^2/2$+additional energy $\epsilon mv_n^2/2$ due to the fact that oscillations occur and that the i.o. is idealized). Due to the assumed smallness of the oscillations ϵ is a small quantity; thus

$$v_\theta\simeq v_n(1+\tfrac{1}{2}\epsilon-\tfrac{1}{2}x'^2-\tfrac{1}{2}y'^2) \tag{4.8}$$

up to terms of the second degree. Equations (4.5) and (4.6) then become, respectively,

$$x''+(1-n)x=F=-(1+\tfrac{1}{2}\epsilon-\tfrac{1}{2}x'^2-\tfrac{1}{2}y'^2)(Z_0+(Z_x-n)x+Z_y y+\tfrac{1}{2}Z_{xx}x^2+\cdots)-y'(Q_0+Q_x x+Q_y y+\cdots)+1+\epsilon+x^2-x'^2-y'^2 \tag{4.9}$$

$$y''+ny=G=x'(Q_0+Q_x x+Q_y y+\cdots)+(1+\tfrac{1}{2}\epsilon-\tfrac{1}{2}x'^2-\tfrac{1}{2}y'^2)(R_0+R_x x+(R_y-n)y+\tfrac{1}{2}R_{xx}x^2+\cdots) \tag{4.10}$$

using some of the results of Section II. It is possible to eliminate the constant terms on the right of Eqs. (4.9) and (4.10) by a slight definition of the i.o. [*viz.*, replacing r_n by $r_n+\epsilon/2(1-n)$ eliminates the constant term in (3.9) to first order, while raising the equilibrium plane by an amount $\sim r_n R_0^0/n$ gets rid of the constant term in (4.10)]. It will be assumed that this has been done, and that only the non-constant terms in F and G need be considered. For the sake of simplicity let

$$a=(1-n)^{\frac{1}{2}},\quad b=(n)^{\frac{1}{2}}. \tag{4.11}$$

If $F\equiv 0$, $G\equiv 0$, equations (4.9) and (4.10) have the solutions[2]

$$x=X\sin(a\theta+\phi) \tag{4.12}$$

$$y=Y\sin(b\theta+\psi), \tag{4.13}$$

X, Y, ϕ, and ψ being suitable constants. Since F and G are not constant, but involve x, y, and θ in a complicated way, the solutions may still be written in the same form, but with X, Y, ϕ, and ψ taken as functions of θ. Imposing the additional conditions

$$x'=aX\cos(a\theta+\phi) \tag{4.14}$$

$$y'=bY\cos(b\theta+\psi), \tag{4.15}$$

‡ The first part of this section is a summary of portions of reference 1 in which a more detailed discussion is given.

[2] See A. Kryloff and N. Bogoliuboff, *Non-Linear Mechanics*, Princeton University Press, 1943 for further details of this method.

one finds that $X(\theta)$, $Y(\theta)$ satisfy the first-order differential equations

$$X' = F\cdot\cos(a\theta+\phi)/a \tag{4.16}$$

$$Y' = G\cdot\cos(b\theta+\psi)/b. \tag{4.17}$$

The right-hand sides of these equations are small, and almost periodic in θ, so that in general, the average of X', Y' over several periods of the motion will be zero or negligible. This will not be the case if there are secular (θ-independent) or almost secular (very long period) terms in the right-hand sides. In this case X' or Y' or both will be of constant sign over many periods of the oscillations, and X, Y may become appreciable despite insignificant initial values. In this approximation, therefore, appreciable oscillations may arise for those values of a and b for which secular terms appear in the right-hand sides of Eqs. (4.16) and (4.17).

In terms of the original equations (4.9) and (4.10) the interpretation of these oscillations is somewhat as follows. The beam has in first approximation two modes of oscillation, a radial mode with frequency $w_r = w_n(1-n)^{\frac{1}{2}}$, and a vertical mode with frequency $w_z = w_n(n)^{\frac{1}{2}}$. The equations of motion are linear in this approximation, and thus the amplitude of the oscillations will be constant and determined by their initial values. In higher approximation, the equations of motion contain complicated non-linear terms (F, G) which may cause suppression of the oscillations due to interference, and will generally prevent any increase in the amplitudes, if not actively decreasing them. For certain particular values of n, however, these non-linear terms give rise to a species of subharmonic resonance, so that the initial oscillations of the beam may increase in amplitude.

It is to be noted that ϕ and ψ also satisfy first-order differential equations, similar in form to (4.16) and (4.17). If one is interested mainly in the amplitudes of individual oscillations, these phase variations will not play a significant role. On the other hand, if the motion is to be traced through from one region of oscillation to another, phase changes will be important, because of their effect on the initial values, and the equations for ϕ' and ψ' may have to be discussed in a similar way to that in which (4.16) and (4.17) are dealt with below. This is, however, not required in the sequel.

V. POSSIBLE OSCILLATIONS

This section is devoted to the task of finding for what values of a and b (i.e., of n) secular terms may arise in $F\cos(a\theta+\phi)$ and $G\cos(b\theta+\psi)$, and to discussing briefly the corresponding oscillations. The calculation is, of course, only carried out up to second-degree terms, and since it is straightforward, though tedious, no details of the first part of the computation are given here.

Since a and b are symmetric about $n=\frac{1}{2}$, and F, G both contain terms of substantially the same type, it is easy to see that any oscillation which occurs for $n=\frac{1}{2}-j$ will also occur for $n=\frac{1}{2}+j$, with the roles of x and y interchanged. Because of this, and also because the beam is usually extracted well before $n=\frac{1}{2}$, only the oscillations in the range $0<n\leqslant\frac{1}{2}$ are discussed here.

Inspection of F and G shows that in this range secular terms may arise from sines or cosines with the following arguments: (i) $(a-2b)\theta$, $n\simeq\frac{1}{5}$; (ii) $(1-3b)\theta$, $n\simeq\frac{1}{9}$; (iii) $(1-2b)\theta$, $n\simeq\frac{1}{4}$; (iv) $(1-2a-b)\theta$, $(2-a-2b)\theta$, $n\simeq 9/25$; (v) $(2-3b)\theta$, $n\simeq 4/9$; (vi) $(a-b)\theta$, $n\simeq\frac{1}{2}$. The corresponding amplitudes X and Y then satisfy equations as follows.

(i) $n\simeq\frac{1}{5}$, $a\simeq 2b$

In this case one finds the coupled oscillations

$$\left.\begin{aligned} X' &= -KY^2\cos p \\ Y' &= 4KXY\cos p \end{aligned}\right\}. \tag{5.1}$$

Here

$$16K = (5)^{\frac{1}{2}}Z_{xx}{}^0 = (5)^{\frac{1}{2}}(6/25-n') \tag{5.2}$$

$p=(a-2b)\theta+u$, and u, the initial phase, depends only on the initial phases of the two individual oscillations, ϕ_0 and ψ_0, $u=\phi_0-2\psi_0$. It follows from (5.1) that $4X^2+Y^2=$constant. These oscillations are dealt with *in extenso* in reference 1, and therefore will not be discussed further here, though similar oscillations are considered in more detail in (iv) below.

(ii) $n\simeq\frac{1}{9}$, $1\simeq 3b$

Here there is only a vertical oscillation, satisfying the equation

$$Y' = |C|\,Y^2\cos p \tag{5.3}$$

$$C = \int_{-\pi}^{\pi}(9Z_{xy}-R_0)e^{i\theta}d\theta/48\pi \tag{5.4}$$

$$p=(1-3b)\theta+u,\ u=-3\psi_0+\text{phase } C.$$

Thus,

$$Y = Y_0/(1-Y_0|C|\cdot\int_0^{\theta}\cos p\cdot d\theta). \tag{5.5}$$

Noting that $(1-3b)\theta\simeq 0$, and choosing the origin of θ at the point where $1=3b$, p may be approximated by

$$p = u-3\theta^2db/d\theta = u-9g^2\theta^2\pi/2 \tag{5.6}$$

$$\pi g^2 = (dn/dr)(dr/d\theta) = n'\cdot\Delta E/4(1-n)E\cdot\pi,$$

i.e.,

$$g^2 = \frac{1}{\pi}\cdot\frac{dn}{dr}\frac{dr}{d\theta} = \frac{n'}{4\pi^2(1-n)}\cdot\frac{\Delta E}{E}. \tag{5.7}$$

ΔE = energy per revolution of the beam, E = total energy. (See Appendix I.)

Thus,

$$\int_0^{\theta}\cos p\cdot d\theta = (\cos u\cdot\mathfrak{C}(3g\theta)+\sin u\cdot\mathfrak{S}(3g\theta))/3g, \tag{5.8}$$

where

$$\mathcal{C}(\theta)=\int_0^\theta \cos\frac{\pi}{2}\theta^2 d\theta, \quad \mathcal{S}(\theta)=\int_0^\theta \sin\frac{\pi}{2}\theta^2\cdot d\theta \quad (5.9)$$

are the usual Fresnel functions.

The behavior of the oscillations thus depends both on the magnitude of C and on the initial phase u. In order to examine the behavior more closely, one assumes that $g^2\theta^2$ is small enough not only for (5.6) to be valid, but also to permit the right-hand side of (5.8) to be replaced by $|C|\cos u\cdot\theta$, so that

$$Y=Y_0/(1-Y_0|C|\cos u\cdot\theta). \quad (5.10)$$

Since Y and $|C|$ are positive, increasing oscillations can only occur for

$$\cos u>0. \quad (5.11)$$

This approximation also defines a limiting value of the path length θ

$$\theta_1=1/Y_0|C|\cos u \quad (5.12)$$

beyond which the method is not valid, despite the smallness of $g^2\theta^2$. A similar, though rather more complicated effect may also take place in the approximation (5.8).

(iii) $n\simeq\frac{1}{4}$, $1\simeq 2b$

Here again there is only a vertical oscillation, the amplitude of which satisfies the equation

$$Y'=|D|Y\cos p \quad (5.13)$$

$$D=\int_{-\pi}^{\pi} Z_x e^{i\theta}d\theta/2\pi \quad (5.14)$$

$$p=(1-2b)\theta+u,\ u=-2\psi_0+\text{phase}(1/D).$$

Thus

$$Y=Y_0\exp(|D|\int_0^\theta \cos p d\theta). \quad (5.15)$$

Proceeding as in (ii)

$$p=u-2\pi g^2\theta^2 \quad (5.16)$$

and hence

$$Y=Y_0\exp(|D|(\cos u\cdot\mathcal{C}(2g\theta)+\sin u\cdot\mathcal{S}(2g\theta))/2g). \quad (5.17)$$

For small enough $g^2\theta^2$, one has

$$Y\simeq Y_0\exp(|D|\cos u\cdot\theta). \quad (5.18)$$

If $\cos u>0$, Y will increase approximately exponentially, with "time constant" $1/w_n|D|\cos u$, and thus, if conditions are favorable, very disturbing oscillations may arise in this region. It is interesting to remark, parenthetically, that there was strong evidence of this type of oscillation in the Princeton cyclotron: the edges of the dees were found to be strongly radioactive in the region $n\simeq\frac{1}{4}$, indicating that the beam was vertically splayed out in this region.

(iv) $n\simeq 9/25$, $2a-b\simeq 1$ and/or $a+2b\simeq 2$

In this case there are coupled oscillations satisfying the equations

$$X'=|L|XY\cos p+|N|Y^2\cos q$$
$$Y'=|M|X^2\cos p^*+|P|XY\cos q^* \quad (5.19)$$

$$L=\int_{-\pi}^{\pi}(5Z_{xy}+3iQ_x)e^{i\theta}d\theta/32\pi$$
$$M=\int_{-\pi}^{\pi}(25R_{xx}+16R_0)e^{i\theta}d\theta/240\pi \quad (5.20)$$

$$N=\int_{-\pi}^{\pi}(25Z_{yy}+30iQ_y+9Z_0)e^{2i\theta}d\theta/320\pi$$
$$P=\int_{-\pi}^{\pi}(5Z_{xx}+4iQ_y)e^{2i\theta}d\theta/24\pi \quad (5.21)$$

$$p=(1-2a+b)\theta+u, \qquad p^*=(1-2a+b)\theta+u^*$$
$$q=(2-a-2b)\theta+s, \qquad q^*=(2-a-2b)\theta+s^*$$
$$u=-2\phi_0+\psi_0+\text{phase}L,\ u^*=-2\phi_0+\psi_0+\text{phase}M$$
$$s=-\phi_0-2\psi_0+\text{phase}N\ \ s^*=-\phi_0-2\psi_0+\text{phase}P$$

The first terms on the right arise from the fundamental Fourier coefficients, and correspond to the near vanishing of $1-2a+b$, while the second terms come from the second harmonics, and correspond to the smallness of $2-a-2b$.

Equations (5.19) are pretty complicated even when the approximation $p=p^*=q=q^*=0$ is made. However, an indication of the sort of thing that happens is given by taking $N=P=0$ (absence of higher harmonics in the field), and placing $u=u^*$. This latter assumption may make it necessary to replace $|M|$ by $-|M|$, to obtain a reasonable representation of what happens if u and u^* differ widely. The equations then become [see (i)]

$$X'=|L|XY\cos p$$
$$Y'=\pm|M|X^2\cos p. \quad (5.22)$$

If the plus sign obtains,

$$(X^2/|L|)-(Y^2/|M|)=k^2, \quad (5.23)$$

and one may put

$$X=k(|L|)^{\frac{1}{2}}\sec(W-W_0),$$
$$Y=k(|M|)^{\frac{1}{2}}\tan(W-W_0) \quad (5.24)$$

whence W satisfies the differential equation

$$W'=-k|L|(|M|)^{\frac{1}{2}}\cos p. \quad (5.25)$$

For small enough θ,

$$W-W_0\simeq -k|L|(|M|)^{\frac{1}{2}}\cos u\cdot\theta. \quad (5.26)$$

Because of (5.24) one again obtains a limiting value for

the angular path length θ

$$\theta_1 = \pi/2k|L|(|M|)^{\frac{1}{2}}\cos u. \tag{5.27}$$

It is clear that in this case both the radial and vertical oscillations may reach uncomfortably large amplitudes, unless their initial values are both zero. This is not so however if the minus sign applies in Eqs. (5.22); the equations then obtained are the analog of (5.1), with the roles of X and Y interchanged. (5.23) is replaced by

$$(X^2/|L|)+(Y^2/|M|)=k^2 \tag{5.28}$$

and one writes

$$X=k(|L|)^{\frac{1}{2}}\sin(W-W_0), \qquad Y=k(|M|)^{\frac{1}{2}}\cos(W-W_0). \tag{5.29}$$

W satisfies the equation

$$W'=k|L|(|M|)^{\frac{1}{2}}\cdot\sin(W-W_0)\cdot\cos p \tag{5.30}$$

so that

$$W-W_0=\arctan\exp\int_0^\theta k|L|(|M|)^{\frac{1}{2}}\cos p\,d\theta \tag{5.31}$$

or, for small θ,

$$W-W_0\simeq\arctan\exp k|L|(|M|)^{\frac{1}{2}}\cos u\cdot\theta. \tag{5.32}$$

Here the combined amplitudes are limited by their initial values, though disturbing oscillations may still occur, due to coupling between the radial and vertical modes, if L and M differ widely in magnitude. Similar results are obtained if there are only second harmonics in the field, *viz.*, $L=M=0$, N, $P\neq 0$.

(v) $n\simeq 4/9$, $2\simeq 3b$

The vertical oscillation has the same form as that occurring at $n\simeq\frac{1}{9}$, except that the coefficient C now depends on the corresponding second harmonics in the field components. [See (ii).]

(vi) $n\simeq\frac{1}{2}$, $a\simeq b$

Using the normalization described following (4.10), the coupled oscillations satisfy the equations

$$\begin{aligned} X' &= BY\cos p \\ Y' &= -BX\cos p \\ B &= Z_y^0/\sqrt{2} \\ p &= (a-b)\theta+u, \quad u=\phi_0-\psi_0. \end{aligned} \tag{5.33}$$

In general $B=0$, since the vertical component of the field is usually symmetric about the plane in which $R_0^0=\int_0^{2\pi}R_0 d\theta/2\pi=0$. The redefinition of coordinates involved in attaining this state of affairs may however distort the median "plane" into slightly conical form (deviation $\sim r_n R_0^0/n$) so that then B is not zero. Equation (5.33) yields the relation

$$X^2+Y^2=k^2B. \tag{5.34}$$

Writing

$$X=k(B)^{\frac{1}{2}}\sin(W-W_0), \quad Y=k(B)^{\frac{1}{2}}\cos(W-W_0) \tag{5.35}$$

one finds

$$W'=B\cos p \tag{5.36}$$

so that, proceeding in the usual way,

$$W-W_0=2^{\frac{1}{2}}B[\cos u\cdot\mathcal{C}(g\theta/2^{\frac{1}{2}})+\sin u\cdot\mathcal{S}(g\theta/2^{\frac{1}{2}})]/g. \tag{5.37}$$

For sufficiently small θ,

$$W-W_0\simeq B\cos u\cdot\theta. \tag{5.38}$$

Thus X and Y are themselves approximately periodic with period

$$2\pi/B\cos u. \tag{5.39}$$

The amplitudes are here again limited by their initial values.

VI. GENERAL REMARKS

The discussion of the previous section was concerned chiefly with the possibility and behavior of self-maintaining or increasing oscillations. It is also fitting to make some brief comments about the oscillations in general, i.e., when the particular cases of Section V do not obtain.

Since in this case the oscillations are small, and (by hypothesis) not self-maintaining, it is sufficient to consider the differential equations (4.9) and (4.10) only up to first-degree terms, *viz.*,

$$x''+a^2x=-(Z_x-n)x-Z_y y-(Z_0-1)-Q_0 y' \tag{6.1}$$

$$y''+b^2y=(R_y-n)y+R_x x+(R_0-R_0^0)+Q_0 x'. \tag{6.2}$$

The coefficients on the right are in general Fourier series in θ with period 2π. Because of this, these equations are closely related to the Mathieu equations: alternatively, attempting to solve these equations by Laplace transforms leads to a set of difference equations. Unless something fairly specific is known about the magnitudes of the Fourier coefficients, there is not much to be gained by pursuing such a mathematically detailed treatment further, all the more in view of the many preceding approximations. Brief consideration will thus only be given to the approximate solution of the system (6.1) and (6.2).

Because of the form of the differential operator on the left, terms containing (say) $\cos k\theta$ on the right will yield similar terms with a multiplicative factor $1/(a^2-k^2)$, $1/(b^2-k^2)$, respectively. Hence only the first few terms need be considered in the Fourier series on the right, at least in an approximate treatment. Incidentally, the values of n for which alarming oscillations may occur may be found in this way (successive substitutions), by placing $x=X\sin a\theta$, $y=Y\sin b\theta$ on the right, and looking for those values of a, b for which some of the a^2-k^2, b^2-k^2 reduce to zero. Second-order terms should also be considered in this case.

It is not difficult to see that even in the absence of an initial deviation there will be a forced oscillation due to

the cylindrical asymmetry of the field, of the form

$$x = -Z_0^1 \cos\theta/n \tag{6.3}$$

$$y = R_0^1 \cos\theta/(1-n^*) \tag{6.4}$$

(if only fundamentals are present in the field). The "oscillations" (6.3), (6.4) denote, respectively, a displacement of the center of the i.o. by an amount $Z_0^1 r_n/n$, and a tilting of its plane by an angle $\sim R_0^1/(1-n)$. If higher harmonics are present the interpretation is no longer so simple: the i.o. is no longer a circle, nor even a plane curve, though it will generally be fairly close to the more idealized i.o. as initially defined. In higher approximation, not only will higher Fourier components appear in x and y but also small constant terms which may be removed by further renormalization of the original i.o. radius r_n, and the i.o. plane [see (4.10)f]; these latter correspond to the fact that any "excess" energy of the beam is shared between the various modes (*viz.*, the rotational, as well as the two vibrational modes).

Inspection of (6.1) and (6.2) shows that if the beam contains some components of its normal modes to start with, it will also contain terms of the type $\sin(k+a)\theta$, etc. In general, if the more exact equations of motion are considered, the deviations from equilibrium have the form

$$\sum_k \sum_l \sum_m A_{klm} \sin[(k+la+mb)\theta+\phi_{klm}] \tag{6.5}$$

so that the motion is generally not even approximately periodic. In special cases when $la+mb$ is an integer, the deviation becomes approximately periodic and much larger; these cases correspond to the oscillations considered in Section V.

APPENDIX I. CHANGE OF RADIUS WITH PATH LENGTH

The radius of the equilibrium orbit given by

$$r = (mc/e)(v/H),$$

H being the vertical component of the field, so that the relation between the increments is

$$(\Delta r/r) = (\Delta v/v) - (\Delta H/H).$$

Since the energy

$$E = mv^2/2$$
$$(\Delta v/v) = (\Delta E/2E),$$

while by definition of n

$$(\Delta H/H) = -n(\Delta r/r).$$

Thus

$$(\Delta r/r) = (\Delta E/2(1-n)E).$$

Taking E to be the increment in energy per revolution (i.e., the energy gained from the r-f field of the dees per revolution) one finds for the average rate of increase of radius per angular path length

$$dr/d\theta = (r \cdot \Delta E/4\pi(1-n)E).$$

APPENDIX II. DESCRIPTION OF THE FIELD

Using the definition (2.6) of the field, and the electromagnetic equations $\mathrm{curl}\mathbf{H}=0$, $\mathrm{div}\mathbf{H}=0$, one has

$$\frac{\partial Z}{\partial\theta} - (1+x)\frac{\partial Q}{\partial y} = 0 \qquad Q + (1+x)\frac{\partial Q}{\partial x} - \frac{\partial R}{\partial\theta} = 0$$

$$\frac{\partial Z}{\partial x} - \frac{\partial R}{\partial y} = 0 \qquad R + (1+x)\frac{\partial R}{\partial x} + \frac{\partial Q}{\partial\theta} + (1+x)\frac{\partial Z}{\partial y} = 0.$$

Inserting the Eqs. (2.7) and (2.8), i.e., the expansions of R, Q, Z, in Taylor series in x, y and Fourier series in θ, and equating coefficients, one finds the following relations, which prove useful in specific calculations:

$$\begin{aligned}
Z_x^k &= R_y^k & ikZ_0^k &= Q_y^k \\
Z_{xy}^k &= R_{yy}^k & ikZ_x^k &= Q_y^k + Q_{xy}^k \\
Z_{xx}^k &= R_{xy}^k & ikZ_y^k &= Q_{yy}^k \\
ikR_0^k &= Q_0^k + Q_x^k & -ikQ_0^k &= R_0^k + R_x^k + Z_y^k \\
ikR_x^k &= Q_{xx}^k + 2Q_x^k & -ikQ_x^k &= R_{xx}^k + 2R_x^k + Z_y^k \\
& & -ikQ_y^k &= R_y^k + R_{xy}^k + Z_{yy}^k.
\end{aligned}$$

Taking $k=0$ these equations show that

$$\int_{-\pi}^{\pi} Q d\theta = 0,$$

which must of course be the case in the absence of a vertical component of the magnetizing current.

The author is indebted to Drs. P. C. Gugelot and H. J. Lipkin for helpful discussions and suggestions.

On the Theory of Dislocations*

JOHN LEE BOGDANOFF
Department of Civil Engineering, Columbia University, New York, New York
(Received May 12, 1950)

If, in a multiply-connected elastic solid, discontinuities are permitted across a stationary barrier in either the strain or its first derivatives or both, dislocations of a more general type than encountered in classical theory are possible. A number of these more general dislocations have been obtained for states of plane and anti-plane strain in a hollow right circular cylinder when the surface of discontinuity is a single stationary plane barrier. Some of the dislocations found possess the characteristic that although the strain is continuous across the barrier the displacement discontinuity is *not* one which would be possible in a rigid body. Examination of the conditions for the uniqueness of solution of the boundary value problems of elasticity reveals that when dislocations of the more general type are admitted appropriate data must be given at each point on the specified barrier in addition to the usual information.

Notation

r, θ, z: Cylindrical coordinates.
$\sigma_\theta, \tau_{r\theta}, \tau_{rz}, \sigma_r, \tau_{\theta z}, \sigma_z$: Components of stress tensor.
$e_{rr}, e_{r\theta}, e_{rz}, e_{\theta\theta}, e_{\theta z}, e_{zz}$: Components of strain tensor.
u_r, u_θ, u_z: Displacement components.
ω_z: z component of rotation.
E: Young's modulus.
σ: Poisson's ratio.

I. INTRODUCTION

In the classical theory of dislocations in a multiply-connected body, Weingarten's theorem[1] plays a fundamental role. According to it, the displacement of the matter immediately on one side of a barrier with respect to the matter on the other side is one which would be possible in a rigid body. Inherent in the proof of this theorem is the assumption that the strain is continuous of class $C^{(2)}$ throughout. We now ask the following question: If discontinuities in the strain or its first derivatives or both are permitted across a stationary barrier, what new displacement discontinuities are allowed? It must be expected in this case that the displacement discontinuity cannot be represented by a displacement possible in a rigid body. The existence of dislocations of a more general type than those of classical theory has been suggested by Goodier.[2]

The six dislocations of classical theory possible in a hollow right circular cylinder are known and have been described in detail by Volterra.[3] In particular there are three in plane strain and one in anti-plane strain. We shall seek dislocations of the less restricted type in the same body for states of plane and anti-plane strain when there is a single plane barrier Ω.

If the strain is discontinuous across Ω, it follows that discontinuities also exist across the same surface in the stress components. It is known[4] however, that in the absence of body forces equilibrium requires the traction on a stationary surface of discontinuity to be continuous across it. Hence, we are restricted to have discontinuities only in those components of the stress which do not enter the expression for the traction on Ω. Because of this condition, it is convenient in what follows to work directly with the stresses.

II. STATEMENT OF PROBLEM

We consider the equilibrium of a hollow right circular cylinder of isotropic elastic material in which there are no body forces. Let the z axis of a cylindrical coordinate system (r, θ, z) coincide with the axis of the cylinder, and let the surface of discontinuity Ω be the intersection of the half-plane $\theta=0$ with the cylinder. The angular coordinate is restricted by

$$0 \leq \theta \leq 2\pi. \tag{2.01}$$

In order to obtain dislocations of the more general type we must seek stress systems in which the stresses or their derivatives or both are discontinuous across Ω. Further, these stresses must satisfy both the differential equations of equilibrium and the compatibility equations; and the traction on Ω must be continuous across this surface.

The procedure followed in obtaining appropriate stress systems is the same for both the anti-plane strain and plane strain problems. First, we assume that at least one of the components of traction on the barrier is expressed by means of functions that are continuous of class $C^{(2)}$ in the simply-connected region formed by introducing Ω and continuous across this surface. Next,

* As used in the present paper, the term dislocation means a state of strain in a multiply-connected elastic solid in which the displacement is discontinuous across one or more barriers. However, this term has a different meaning in theories concerning the plasticity of matter. For a definition of this second meaning and for a discussion of the uses of the dislocations defined above in plastic theories, the reader should consult the references given at the end of Chapter 2 in the book, *The Inelastic Behavior of Engineering Materials and Structures*, A. M. Freudenthal (John Wiley and Sons, Inc., New York, 1950).

[1] A. E. H. Love, *Mathematical Theory of Elasticity* (Cambridge University Press, London 1927), 4th Edition, pp. 221–224.

[2] J. N. Goodier, "On the theory of dislocations and on application to the plane stress problem of a system of forces acting at a hole," 5th International Congress for Applied Mechanics (John Wiley & Sons, Inc., New York, 1938), pp. 129–133.

[3] V. Volterra, "Sur l'equilibre des corps elastiques multiplement connexes," Paris, Ann. Ec. norm. (Ser. 3), t 24, 1907, pp. 401–517.

[4] J. W. Gibbs, "On the equilibrium of hetrogeneous substances," *Collected Works of J. W. Gibbs* (Yale University Press, New Haven, 1948), Vol. 1, pp. 183–194.

using this assumption, stresses are derived (in which discontinuities may exist) that satisfy the differential equations of equilibrium and compatibility relations. Finally, after rejecting those systems that either are continuous of class $C^{(2)}$ throughout the cylinder or have discontinuities which do not meet the requirements of continuous traction on crossing Ω, the displacements are determined. Any remaining stress systems that have discontinuities in displacement across the barrier, constitute dislocations of the desired type.

The barrier is at $\theta=0$ in what follows. To secure complete freedom in choosing the angular position, corresponding dislocations are needed also for Ω at $\theta=\pi/2$. Since the procedures are identical for both positions, we shall indicate only the initial step when the barrier is in the second position. It will be recalled that, in the case of dislocations in classical theory, the position of the barrier is immaterial.

3. ANTI-PLANE STRAIN

In a state of plane strain, the stress components, σ_r, σ_θ, σ_z and $\tau_{r\theta}$ are functions of r and θ while τ_{rz} and $\tau_{\theta z}$ vanish. In a state of anti-plane strain, the converse is true. Then, the differential equations of equilibrium require the existence of a function $\varphi(r,\theta)$ such that

$$\tau_{rz}=r^{-1}\partial\varphi/\partial\theta, \quad \tau_{\theta z}=-\partial\varphi/\partial r; \tag{3.01}$$

and the compatibility relations require

$$(\partial^2/\partial r^2+r^{-1}\partial/\partial r+r^{-2}\partial^2/\partial\theta^2)\varphi\equiv\nabla^2\varphi=d_0\,(\text{const.}), \tag{3.02}$$

where φ is a continuous function of class $C^{(2)}$ in the simply-connected cylinder. For continuity of traction on Ω, we must have

$$\Delta\tau_{\theta z}\equiv\tau_{\theta z}\,|_{\theta=2\pi}-\tau_{\theta z}\,|_{\theta=0}=0. \tag{3.03}$$

Let us now assume that $\tau_{\theta z}$ is expressed by

$$-(\tau_{\theta z})_n=\xi_n(r)\theta\sin n\theta+\eta_n(r)\cos n\theta+\eta_0(r) \tag{3.04}$$

where n is an integer, $\xi_n(r)$, $\eta_n(r)$, and $\eta_0(r)$ are continuous functions of class $C^{(2)}$, and

$$0\leq\theta\leq 2\pi.$$

Equation (3.03) is satisfied by this assumption.

The stress function φ_n we seek must satisfy (3.02) and

$$\partial\varphi_n/\partial r=\xi_n(r)\theta\sin n\theta+\eta_n(r)\cos n\theta+\eta_0(r). \tag{3.05}$$

The function φ which satisfies these equations is

$$\varphi\equiv\sum_{n=1}^{\infty}\varphi_n=\sum_{n=1}^{\infty}a_nr^n(\theta\sin n\theta-\log r\cos n\theta)+\sum_{n=1}^{\infty}b_nr^{-n}(\theta\sin n\theta+\log r\cos n\theta)$$
$$+L(\theta^2-\log^2 r)+\sum_{n=1}^{\infty}(c_nr^n+d_nr^{-n})\cos n\theta+a_0+b_0\log r+c_0\theta+\frac{d_0r^2}{4}. \tag{3.06}$$

The terms on the last line of this equation give stresses that are continuous of class $C^{(2)}$ throughout the cylinder. Hence, they are of no interest to us, except for noting that the term $b_0\log r$ represents a dislocation of classical theory of order 3.[8] The stresses for the remaining terms are

$$\left.\begin{aligned}\tau_{rz}&=\sum_{n=1}^{\infty}a_nr^{n-1}[n\theta\cos n\theta+(n\log r+1)\sin n\theta]\\&\quad+\sum_{n=1}^{\infty}b_nr^{-n-1}[n\theta\cos n\theta\\&\quad-(n\log r-1)\sin n\theta]+L\theta r^{-1},\\\tau_{\theta z}&=-\sum_{n=1}^{\infty}a_nr^{n-1}[n\theta\sin n\theta-(n\log r+1)\cos n\theta]\\&\quad+\sum_{n=1}^{\infty}b_nr^{-n-1}[n\theta\sin n\theta\\&\quad+(n\log r-1)\cos n\theta]+Lr^{-1}\log r.\end{aligned}\right\} \tag{3.07}$$

The components of displacement u_r, u_θ, u_z are related to the strains e_{rz}, $e_{\theta z}$ and the stresses τ_{rz}, $\tau_{\theta z}$ by the following expressions

$$\left.\begin{aligned}2\tau_{rz}(1+\sigma)/E&=e_{rz}=\partial u_r/\partial z+\partial u_z/\partial r,\\2\tau_{\theta z}(1+\sigma)/E&=e_{\theta z}=r^{-1}\partial u_z/\partial\theta+\partial u_\theta/\partial z,\end{aligned}\right\} \tag{3.08}$$

where σ is Poisson's ratio and E is Young's modulus. The displacements corresponding to (3.07) are

$$\left.\begin{aligned}u_r&=0,\\u_\theta&=0,\\u_z&=2(1+\sigma)E^{-1}\{\sum_{n=1}^{\infty}a_nr^n(\theta\cos n\theta+\log r\sin n\theta)\\&\quad-\sum_{n=1}^{\infty}b_nr^{-n}(\theta\cos n\theta-\log r\sin n\theta)+L\theta\log r\},\end{aligned}\right\} \tag{3.09}$$

the terms which represent the displacement of the cylinder as a rigid body being omitted.

Inspection of (3.07) and (3.09) shows that τ_{rz} and u_z are discontinuous across Ω, and $\tau_{\theta z}$, u_r, and u_θ are not. Therefore, (3.07) represent dislocations of the desired

type. The discontinuities have the forms

$$\left.\begin{aligned}\Delta\tau_{rz}&\equiv\tau_{rz}\big|_{\theta=2\pi}-\tau_{rz}\big|_{\theta=0}\\&=2\pi\Big[\sum_{n=1}^{\infty}n(a_nr^{n-1}+b_nr^{-n-1})+Lr^{-1}\Big],\\\Delta u_z&\equiv u_z\big|_{\theta=2\pi}-u_z\big|_{\theta=0}=4\pi(1+\sigma)E^{-1}\\&\quad\times\Big[\sum_{n=1}^{\infty}(a_nr^n-b_nr^{-n})+L\log r\Big].\end{aligned}\right\}\tag{3.10}$$

We also note that the stress system given by $L(\theta^2-\log^2 r)$ has a resultant force acting on a cylindrical surface which encloses the z axis.

Equation (3.04) may be altered to place Ω at $\theta=\pi/2$ by assuming

$$-(\tau_{\theta z})_n=\xi_n(r)\theta\cos n\theta+\eta_n(r)\sin n\theta+\eta_0(r).\tag{3.11}$$

4. PLANE STRAIN

We assume that the displacements u_r and u_θ are functions of r and θ only while u_z is zero. The strain components e_{rz}, $e_{\theta z}$, e_{zz} vanish throughout the cylinder; and the formulas connecting the strains, displacements and stresses are

$$\left.\begin{aligned}e_{rr}&=\partial u_r/\partial r=(\sigma_r-\sigma'\sigma_\theta)/E',\\e_{r\theta}&=\partial u_\theta/\partial r-r^{-1}u_\theta+r^{-1}\partial u_r/\partial\theta\\&=2(1+\sigma')\tau_{r\theta}/E',\\e_{\theta\theta}&=r^{-1}\partial u_\theta/\partial\theta+r^{-1}u_r=(\sigma_\theta-\sigma'\sigma_r)/E',\end{aligned}\right\}\tag{4.01}$$

and

$$\sigma_z=\sigma'(\sigma_r+\sigma_\theta)\tag{4.02}$$

where $\sigma'=\sigma/1-\sigma$, $E'=E/1-\sigma^2$. Then the differential equations of equilibrium are the conditions for the existence of a function $\chi(r,\theta)$ such that

$$\sigma_r=\frac{\partial\chi}{r\partial r}+\frac{\partial^2\chi}{r^2\partial\theta^2},\quad\tau_{r\theta}=-\frac{\partial}{\partial r}\left(\frac{\partial\chi}{r\partial\theta}\right),\quad\sigma_\theta=\frac{\partial^2\chi}{\partial r^2},\tag{4.03}$$

where Airy's stress function χ is a continuous function of class $C^{(2)}$ in the simply-connected region. The compatibility relations require that χ be biharmonic:

$$\nabla^4\chi=0.\tag{4.04}$$

The requirement of continuity of traction on crossing Ω is expressed by

$$\left.\begin{aligned}\Delta\tau_{r\theta}&\equiv\tau_{r\theta}\big|_{\theta=2\pi}-\tau_{r\theta}\big|_{\theta=0}=0,\\\Delta\sigma_\theta&\equiv\sigma_\theta\big|_{\theta=2\pi}-\sigma_\theta\big|_{\theta=0}=0.\end{aligned}\right\}\tag{4.05}$$

We assume first that $\tau_{r\theta}$ and σ_θ are given by the real and imaginary parts of expressions of the type

$$(\tau_{r\theta})_n=f_n(r)e^{in\theta},\quad(\sigma_\theta)_n=g_n(r)e^{in\theta},\tag{4.06}$$

where $f_n(r)$ and $g_n(r)$ are continuous functions of class $C^{(2)}$ and n is an integer including zero. Clearly, these expressions satisfy (4.05).

We must now determine stresses $(\sigma_r)_n$, $(\tau_{r\theta})_n$, and $(\sigma_\theta)_n$ that satisfy the differential equations of equilibrium, the compatibility equations, and (4.06); in the process, we must not exclude those stress systems for which $(\sigma_r)_n$ is discontinuous. These requirements are fulfilled provided the stresses are derived from an Airy stress function χ_n which satisfies

$$-\frac{\partial}{\partial r}\left(\frac{\partial\chi_n}{r\partial\theta}\right)=f_n(r)e^{in\theta},\quad\frac{\partial^2\chi_n}{\partial r^2}=g_n(r)e^{in\theta},$$

$$\nabla^4\chi_n=0.\tag{4.07}$$

Upon integrating these equations, we obtain

$$\begin{aligned}\chi=\sum_{n=1}^{\infty}\chi_n&=A_0r^2+B_0r^2(\log r-1)+C_0\log r+D_0\theta\\&\quad+(A_1r+B_1r^{-1}+B\theta r+C_1r^3+D_1r\log r)e^{i\theta}\\&\quad+\sum_{n=2}^{\infty}(A_nr^n+B_nr^{-n}+C_nr^{n+2}+D_nr^{-n+2})e^{in\theta}\\&\quad+L(\theta^2-\log^2 r)re^{i\theta}.\end{aligned}\tag{4.08}$$

The stresses corresponding to the terms of the first three lines of (4.08) are continuous of class $C^{(2)}$ throughout the cylinder.[5] Therefore, the only function of immediate interest is

$$L(\theta^2-\log^2 r)re^{i\theta}.\tag{4.09}$$

We note in passing that the classical dislocations of orders 1, 2 and 6[5] are given by $Dr\log r\sin\theta$, $D'r\log r\cos\theta$, $B_0r^2\log r$.

The stresses for the real and imaginary parts of (4.09) are respectively

$$\left.\begin{aligned}\sigma_r&=2L_2r^{-1}[(1-\log r)\cos\theta-2\theta\sin\theta],\\\tau_{r\theta}&=-2L_2r^{-1}\log r\sin\theta,\\\sigma_\theta&=-2L_2r^{-1}(1+\log r)\cos\theta,\end{aligned}\right\}\tag{4.10}$$

and

$$\left.\begin{aligned}\sigma_r&=2L_1r^{-1}[(1-\log r)\sin\theta+2\theta\cos\theta],\\\tau_{r\theta}&=2L_1r^{-1}\log r\cos\theta,\\\sigma_\theta&=-2L_1r^{-1}(1+\log r)\sin\theta.\end{aligned}\right\}\tag{4.11}$$

Examination of these formulas reveals that in (4.10) all the stresses (and hence the strains) are continuous across Ω, but $\partial\sigma_r/\partial\theta$ is discontinuous; while in (4.11), σ_r and $\partial\sigma_r/\partial r$ are discontinuous across the same surface, the discontinuity in σ_r being

$$\Delta\sigma_r\equiv\sigma_r\big|_{\theta=2\pi}-\sigma_r\big|_{\theta=0}=8\pi L_1r^{-1}.\tag{4.12}$$

Both stress systems have a resultant force acting on a cylindrical surface enclosing the z axis.

The displacements corresponding to (4.10) and (4.11)

[5] These three lines constitute the stress function given by J. H. Michell in "On the direct determination of stress in an elastic solid," Proc. London Math. Soc. 31, 100–124 (1900).

are given by

$$\left.\begin{aligned} E'u_r &= L_2\{[2(1+\sigma')\log r-(1-\sigma')\log^2 r]\cos\theta \\ &\quad +[(1-\sigma')\theta^2-4\theta\log r]\sin\theta\}, \\ E'u_\theta &= L_2\{[(1-\sigma')\theta^2-4\sigma'\theta-2(1-\sigma') \\ &\quad -4\theta\log r]\cos\theta+(1-\sigma') \\ &\quad \times[\log^2 r-2(1+\theta)]\sin\theta\}, \end{aligned}\right\} \quad (4.13)$$

and

$$\left.\begin{aligned} E'u_r &= L_1\{[4\theta\log r-(1-\sigma')\theta^2]\cos\theta \\ &\quad +[2(1+\sigma')\log r-(1-\sigma')\log^2 r]\sin\theta\}, \\ E'u_\theta &= L_1\{(1-\sigma')[2(1+\theta)-\log^2 r]\cos\theta \\ &\quad +[(1-\sigma')\theta^2-4\sigma'\theta-2(1-\sigma') \\ &\quad -4\theta\log r]\sin\theta\}, \end{aligned}\right\} \quad (4.14)$$

where the rigid body terms have been omitted. The discontinuities in displacement across Ω are

$$\Delta u_r=\text{const.},\quad \Delta u_\theta=\text{const.}-8\pi L_2\log r/E', \quad (4.15)$$

$$\Delta u_r=\text{const.}+8\pi L_1\log r/E',\quad \Delta u_\theta=\text{const.}, \quad (4.16)$$

respectively, where $\Delta u_r\equiv u_r\big|_{\theta=2\pi}-u_r\big|_{\theta=0}$ and $\Delta u_\theta\equiv u_\theta\big|_{\theta=2\pi}-u_\theta\big|_{\theta=0}$. Since the displacements have discontinuities on crossing Ω involving $\log r$ the stress systems (4.10) and (4.11) represent dislocations of the general type.

When Ω is at $\theta=\pi/2$, the roles of the real and imaginary parts of (4.09) are interchanged.

The stress function (4.09) was obtained on the assumption that $\tau_{r\theta}$ and σ_θ have the forms given by (4.06); obviously this is not the only starting point. However, rather than assume other forms for these stress components and proceed as before, it is convenient at this point to make use of information already obtained. We recall that both $\chi_1=r\varphi_1\sin\theta+\varphi_2$ and $\chi_2=r\varphi_1\cos\theta+\varphi_2$, where φ_1 and φ_2 are harmonic functions, are solutions of $\nabla^4\chi=0$. For φ_1 and φ_2 we choose the function given by the first two lines of (3.06). Thus,

$$\begin{aligned} \chi_1 &= \sum_{n=1}^{\infty}[A_n r^{n+1}(\theta\sin n\theta-\log r\cos n\theta) \\ &\quad +B_n r^{-n+1}(\theta\sin n\theta+\log r\cos n\theta)]\sin\theta \\ &\quad +\sum_{n=1}^{\infty}[C_n r^n(\theta\sin n\theta-\log r\cos n\theta) \\ &\quad +D_n r^{-n}(\theta\sin n\theta+\log r\cos n\theta)] \\ &\quad +L_1(\theta^2-\log^2 r)r\sin\theta+G_1(\theta^2-\log^2 r) \end{aligned} \quad (4.17)$$

and

$$\begin{aligned} \chi_2 &= \sum_{n=1}^{\infty}[A_n' r^{n+1}(\theta\sin n\theta-\log r\cos n\theta) \\ &\quad +B_n' r^{-n+1}(\theta\sin n\theta+\log r\cos n\theta)]\cos\theta \\ &\quad +\sum_{n=1}^{\infty}[C_n' r^n(\theta\sin n\theta-\log r\cos n\theta) \\ &\quad +D_n' r^{-n}(\theta\sin n\theta+\log r\cos n\theta)] \\ &\quad +L_2(\theta^2-\log^2 r)r\cos\theta+G_2(\theta^2-\log^2 r). \end{aligned} \quad (4.18)$$

The requirement that the traction on Ω be continuous across this surface can be met by setting

$$C_n=D_n=G_1=0,\quad (n=1, 2, \cdots); \quad (4.19)$$

and

$$\left.\begin{aligned} C_n' &= -A_{n-1}'\frac{n-1}{n}, \quad (n=2, 3, \cdots), \\ D_n' &= -B_{n+1}\frac{n+1}{n}, \quad (n=1, 2, \cdots), \\ G_2 &= -B_1. \end{aligned}\right\} \quad (4.20)$$

With these equations, we find that the stresses corresponding to (4.17) and (4.18) are

$$\left.\begin{aligned} 2\sigma_r &= \sum_{n=1}^{\infty}A_n r^{n-1}\{-(2n-3)\sin(n-1)\theta+(2n+1)\sin(n+1)\theta-n(n-3)\theta\cos(n-1)\theta+n(n+1)\theta\cos(n+1)\theta \\ &\quad -\log r[-n(n-3)\sin(n-1)\theta+n(n+1)\sin(n+1)\theta]\}+\sum_{n=1}^{\infty}B_n r^{-n-1}\{-(2n-1)\sin(n-1)\theta \\ &\quad +(2n+3)\sin(n+1)\theta-n(n-1)\theta\cos(n-1)\theta+n(n+3)\theta\cos(n+1)\theta+\log r[n(n-1)\sin(n-1)\theta \\ &\quad -n(n+3)\sin(n+1)\theta]\}+4L_1 r^{-1}[(1-\log r)\sin\theta+2\theta\cos\theta], \\ 2\tau_{r\theta} &= \sum_{n=1}^{\infty}A_n r^{n-1}\{(2n-1)\cos(n-1)\theta-(2n+1)\cos(n+1)\theta-n(n-1)\theta\sin(n-1)\theta+n(n+1)\theta\sin(n+1)\theta \\ &\quad +\log r[n(n-1)\cos(n-1)\theta-n(n+1)\cos(n+1)\theta]\}+\sum_{n=1}^{\infty}B_n r^{-n-1}\{(2n-1)\cos(n-1)\theta \\ &\quad +(2n+1)\cos(n+1)\theta-n(n-1)\theta\sin(n-1)\theta+n(n+1)\theta\sin(n+1)\theta \\ &\quad +\log r[-n(n-1)\cos(n-1)\theta+n(n+1)\cos(n+1)\theta]\}+4L_1 r^{-1}(1+\log r)\sin\theta, \\ 2\sigma_\theta &= \sum_{n=1}^{\infty}A_n r^{n-1}\{(2n+1)[\sin(n-1)\theta-\sin(n+1)\theta]+n(n+1)\theta[\cos(n-1)\theta \\ &\quad -\cos(n+1)\theta]+n(n+1)\log r[\cos(n-1)\theta-\cos(n+1)\theta]\} \\ &\quad +\sum_{n=1}^{\infty}B_n r^{-n-1}\{(2n-1)[\sin(n-1)\theta-\sin(n+1)\theta]+n(n-1)\theta[\cos(n-1)\theta-\cos(n+1)\theta] \\ &\quad -n(n-1)\log r[\sin(n-1)\theta-\sin(n+1)\theta]\}-4L_1 r^{-1}(1+\log r)\sin\theta; \end{aligned}\right\} \quad (4.21)$$

and

$$
\left.
\begin{aligned}
2\sigma_r = &\sum_{n=1}^{\infty} A_n' r^{n-1}\Big\{(2n-3)\cos(n-1)\theta - \frac{(n-1)(2n+1)}{n+1}\cos(n+1)\theta - n(n-3)\theta\sin(n-1)\theta \\
&+ n(n-1)\theta\sin(n+1)\theta - \log r[n(n-3)\cos(n-1)\theta - n(n-1)\cos(n+1)\theta]\Big\} \\
&+ \sum_{n=2}^{\infty} B_n' r^{-n-1}\Big\{-\frac{(n+1)(2n-1)}{n-1}\cos(n-1)\theta + (2n+3)\cos(n+1)\theta + n(n+1)\theta\sin(n-1)\theta \\
&- n(n+3)\theta\sin(n+1)\theta + \log r[n(n+1)\cos(n-1)\theta - n(n+3)\cos(n+1)\theta]\Big\} \\
&+ B_1' r^{-2}[5\cos 2\theta - 1 - 4\theta\sin\theta + 2\log r(1-2\cos 2\theta)] + 4L_2 r^{-1}[\cos\theta - 2\theta\sin\theta - \log r\cos\theta], \\
2\tau_{r\theta} = &-\sum_{n=1}^{\infty} A_n' r^{n-1}\Big\{(2n-1)\sin(n-1)\theta - \frac{(n-1)(2n+1)}{n+1}\sin(n+1)\theta + n(n-1)\theta[\cos(n-1)\theta - \cos(n+1)\theta] \\
&+ n(n-1)\log r[\sin(n-1)\theta - \sin(n+1)\theta]\Big\} - \sum_{n=1}^{\infty} B_n' r^{-n-1}\Big\{\frac{(n+1)(2n-1)}{n-1}\sin(n-1)\theta \\
&- (2n+1)\sin(n+1)\theta + n(n+1)\theta[\cos(n-1)\theta - \cos(n+1)\theta] - n(n+1)\log r[\sin(n-1)\theta \\
&- \sin(n+1)\theta]\Big\} + B_1' r^{-2}[\sin 2\theta - 2\theta(1-\cos 2\theta) + 2(1-\log r)\sin 2\theta] - 4L_2 r^{-1}\sin\theta, \\
2\sigma_\theta = &\sum_{n=1}^{\infty} A_n' r^{n-1}\Big\{-(2n+1)\cos(n-1)\theta - \frac{(n-1)(2n+1)}{n+1}\cos(n+1)\theta + n(n+1)\theta\sin(n-1)\theta \\
&- n(n-1)\theta\sin(n+1)\theta - \log r[n(n+1)\cos(n-1)\theta - n(n-1)\cos(n+1)\theta]\Big\} \\
&+ \sum_{n=2}^{\infty} B_n' r^{-n-1}\Big\{\frac{(n+1)(2n-1)}{n-1}\cos(n-1)\theta - (2n-1)\cos(n+1)\theta - n(n+1)\theta\sin(n-1)\theta \\
&+ n(n-1)\theta\sin(n+1)\theta - \log r[n(n+1)\cos(n-1)\theta - n(n-1)\cos(n+1)\theta]\Big\} \\
&+ B_1' r^{-2}[1 - \cos 2\theta - 2\log r] - 4L_2 r^{-1}(1+\log r)\cos\theta,
\end{aligned}
\right\} \quad (4.22)
$$

respectively. By means of (4.01), the displacements can be found for these stress systems. The discontinuities in the stresses and the displacements across Ω are

$$
\left.
\begin{aligned}
&\Delta\sigma_r = 4\pi[\sum_{n=1}^{\infty} n(A_n r^{n-1} + B_n r^{-n-1}) + 2L_1 r^{-1}], \\
&E'\Delta u_r = 4\pi[\sum_{n=1}^{\infty}(A_n r^n - B_n r^{-n}) + 2L_1\log r], \\
&\Delta\sigma_\theta = \Delta\tau_{r\theta} = \Delta u_\theta = 0;
\end{aligned}
\right\} \quad (4.23)
$$

and

$$
\left.
\begin{aligned}
E'\Delta u_\theta = 4\pi[&-\sum_{n=1}^{\infty} A_n' r^n + \sum_{n=2}^{\infty} B_n' r^{-n} \\
&+ B_1' r^{-1} + 2L_2\log r], \\
&\Delta u_r = \Delta\sigma_r = \Delta\tau_{r\theta} = \Delta\sigma_\theta = 0.
\end{aligned}
\right\} \quad (4.24)
$$

Equations (4.21) and (4.22) thus represent general dislocations.

As before, χ_1 and χ_2 may be suitably altered to place the barrier at $\theta = \pi/2$.

5. DISCUSSION

After the first draft of this paper was prepared, Mann[6] published an article in which many of the stress functions obtained above were given. In addition, she also solved several boundary value problems (of interest in connection with slip theories of plasticity) for states of plane and anti-plane strain in a hollow cylinder, and she indicated a general procedure for dealing with such problems. There is no need then for additional discussion of these aspects of the solutions.

It was not noticed in reference 6, however, that it is possible to divide the dislocation in the plane strain problem into two distinct classes, and those stress functions involving $(\theta^2 - \log^2 r)$, i.e.

$$L_2(\theta^2 - \log^2 r)r\cos\theta, \quad L_1(\theta^2 - \log^2 r)r\sin\theta, \quad (4.09a)$$

$$L(\theta^2 - \log^2 r) \quad (3.06a)$$

were not found. As these two points are of some importance, we shall comment on them and then conclude with a remark on the requirements for uniqueness of solution of boundary value problems involving a sta-

[6] E. H. Mann, "An elastic theory of dislocations," Proc. Royal Soc. A199, No. 1058, 376-394 (1949).

tionary barrier across which the strains or their derivatives are discontinuous.

In a course of the proof of Weingarten's theorem, we find by means of Cesaro's theorem[7] that for the plane strain problem

$$\left.\begin{aligned}\Delta u_r^{(1)} &= \Delta u_r^{(0)} + \int_{r_0}^{r_1} \Delta e_{rr} dr,\\ \Delta u_\theta^{(1)} &= \Delta u_\theta^{(0)} + (r_1 - r_0)\Delta\omega_z^{(0)} + \int_{r_0}^{r_1} \{\Delta e_{r\theta}/2\\ &\quad - (r_1 - r)[\Delta\partial e_{r\theta}/2\partial r - \Delta(\partial e_{rr}/r\partial\theta)]\} dr\end{aligned}\right\} \quad (5.01)$$

where it has been assumed that the strains are continuous of class $C^{(2)}$ and satisfy the compatibility relations in the simply-connected cylinder, r_0 and r_1 are any two points on Ω, and ω_z is the z component of the rotation. As the traction on Ω must be continuous across this surface,

$$\Delta e_{r\theta} = \Delta(\partial e_{r\theta}/\partial r) = 0, \quad (5.02)$$

for an isotropic solid. Hence, (5.01) becomes

$$\left.\begin{aligned}\Delta u_r^{(1)} &= \Delta u_r^{(0)} + \int_{r_0}^{r_1} \Delta e_{rr} dr,\\ \Delta u_\theta^{(1)} &= \Delta u_\theta^{(0)} + (r_1 - r_0)\Delta\omega_z^{(0)}\\ &\quad + \int_{r_0}^{r_1} (r_1 - r)\Delta(\partial e_{rr}/r\partial\theta) dr.\end{aligned}\right\} \quad (5.03)$$

In the classical dislocation theory, the strains are continuous of class $C^{(2)}$ throughout the body. The integrals on the right-hand side of (5.03) therefore vanish, and the displacement discontinuity has the form predicted by Weingarten's theorem.

Two possibilities arise in the case of dislocations of the general type. First, if we assume that the strains (and hence the stresses) are continuous across Ω, $\Delta e_{rr} = 0$, but discontinuities in displacement can still exist provided that $\Delta(\partial e_{rr}/\partial\theta) = 0$; in this case

$$\begin{aligned}\Delta u_r^{(1)} &= \Delta u_r^{(0)},\\ \Delta u_\theta^{(1)} &= \Delta u_\theta^{(0)} + (r_1 - r_0)\Delta\omega_z^{(0)}\\ &\quad + \int_{r_0}^{r_1} (r_1 - r)\Delta(\partial e_{rr}/r\partial\theta) dr.\end{aligned} \quad (5.04)$$

Second, if we assume that $\Delta e_{rr} \neq 0$ and $\Delta(\partial e_{rr}/\partial\theta) = 0$, the displacement discontinuity across Ω is

$$\begin{aligned}\Delta u_r^{(1)} &= \Delta u_r^{(0)} + \int_{r_0}^{r_1} \Delta e_{rr} dr,\\ \Delta u_\theta^{(1)} &= \Delta u_\theta^{(0)} + (r_1 - r_0)\Delta\omega_z^{(0)}.\end{aligned} \quad (5.05)$$

Thus, there is a division of the general dislocations belonging to the plane strain problem considered in this paper into two classes. In the class given by (5.04), the strains and u_r are continuous across Ω and u_θ is discontinuous; and Δe_{rr}, $\Delta e_{\theta\theta}$, and Δu_r are not zero while $\Delta e_{r\theta}$ and Δu_θ are zero in the class expressed by (5.05), assuming $\Delta u_r^{(0)} = \Delta u_\theta^{(0)} = \Delta\omega_z^{(0)} = 0$, of course.

The dislocations belonging to the class represented by (5.05), i.e. (4.22), reveal an omission in a proof of Weingarten's theorem given in reference 2. In the course of this proof, it is stated that if the strains are continuous across a barrier, which is a surface of discontinuity in displacement, the only possible displacement discontinuity is one that might take place in a rigid body. The stresses, and hence the strains, are continuous across Ω in the dislocations of (4.22), but the displacement discontinuity cannot be represented by the displacement of a rigid body. Hence, we must assume that the strains, together with at least their first derivatives, are continuous across a barrier in order to establish the theorem.

Equations (4.21) represent all the dislocations of the two classes needed to represent specified discontinuities in u_θ and u_r, respectively. For, by assigning appropriate values to the coefficients in these equations, it is seen, from (4.23) and (4.24), that we can express any given discontinuity in u_θ and u_r which is an analytic function of r every where except for possible singularities at $r = 0$ and $r = \infty$. The importance of having the terms of the form (4.09a) is that they are required in order to provide the logrithmic singularity in displacement discontinuity at the axis of the cylinder.

When the solid is not isotropic, the division into two classes of the dislocations in this plane strain problem is not usually possible.

In the anti-plane strain problem considered, there is only one class of general dislocations. The second of equations (3.10) shows that (3.07) provides all dislocations required to represent any specified discontinuity in u_z. The term (3.06a) provides the logrithmic singularity in Δu_z at $r = 0$.

When general dislocations are admitted, the theorems on uniqueness of solution require modification. In addition to the usual quantities, we must now specify at each point on Ω either the traction or the displacement discontinuity or an appropriate combination of the components of each. With the dislocations of classical theory, we recall that we only need specify either the resultant force and couple acting on the barrier or the six constants of the displacement discontinuity or a suitable combination of these quantities. The location and shape of the barrier must also be given when dealing with general dislocations, whereas this information does not have to be supplied in classical theory.

The author wishes to acknowledge that the stress functions given in Eqs. (3.06a) and (4.09a) were disclosed to him by R. D. Mindlin. The author also wishes to thank Prof. Mindlin for helpful discussions during the course of this study.

[7] E. Cesaro, see references 1 and 3.

A Microwave Study of the High Pressure Arc

J. D. Cobine, E. P. Cleary,* and W. C. Gray†
General Electric Research Laboratory, Schenectady, New York
(Received April 26, 1950)

The impedance of an atmospheric pressure d.c. arc was measured at 1000 mc/sec. by means of a coaxial line and standing wave detector. The reactance and resistance of the arc, an essentially lumped load at the end of the line, increase with length of arc. The resistance decreases with increasing current and is approximately the same as the d.c. "resistance." The reactance is capacitive and nearly independent of current. Air, argon, and helium were studied at currents from 1–4 amp. d.c. arc current.

INTRODUCTION

THE following research was initiated in order to determine the possibilities of using microwave techniques to investigate high pressure heavy-current arcs. Measurements made with microwaves would be expected to indicate the actual electron conditions, since only electrons can respond to such high frequencies. The d.c. arc was studied as a lumped load at the end of a coaxial line. An r-f generator and measuring equipment with a 10-cm wave-length was used to study the discharge. The use of cavities and a 3-cm wave-length equipment was found to be unsatisfactory because of the high value of electron density in the high pressure arc.

APPARATUS

The arc chamber, shown schematically in Fig. 1, was arranged so that the arc burned as a section of the inner conductor of the coaxial line. Because of the necessity for water-cooled tungsten electrodes and in order to confine the position of the arc foot-points, the arc was placed ½ wave-length from the shorted end of the line to produce the effect of a short circuit at one electrode surface. It was believed that by keeping the arc length less than 0.1λ the condition of a pure lumped load would be rather closely approached. The shorted end assembly permits changes in the arc length. The arc chamber was mounted vertically and a stabilizing flow of gas was provided by means of three gas inlets set to give a spiral flow of gas around the arc column and exhaust through outlet ports at the top of the device. The fixed electrode assembly had a quarter-wave choke stub to permit direct connection to the d.c. and to admit cooling water. A tuned choke was used with a tapered adapter to connect by a flexible conductor to the microwave generator.

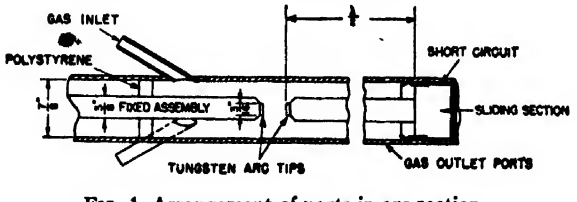


FIG. 1. Arrangement of parts in arc section.

A portion of the outer conductor below the arc chamber consisted of a slotted section for standing wave measurements.

The experimental arrangement is shown in Fig. 2. The microwave generator was an AN/APT5 Radar Set (Carpet Jamming Transmitter) set at 30 cm wave-length, unmodulated, and at minimum power output (~0.1 watt). The attenuator was a length of lossy line having an attenuation at 1000 mc/sec. of 11 db. The arc was maintained by the laboratory d.c. supply with a ballast inductance L and a current controlling resistance R in series. The standing wave measurements were made by means of a crystal detector and microammeter with the slotted line section.

EXPERIMENTAL RESULTS

The impedance, reactance, and resistance of the arc at a 30 cm wave-length were determined by calculation from standing wave data in the usual way with due regard for the reactance of the gap without the arc. It was found that each of these quantities increase with arc length and the resistance decreases with increasing d.c. current in the range 1 to 4 amperes. The arc reactance was found to be capacitive.

Figure 3 shows the resistance of arcs in air for currents from 1 to 4 amperes. A d.c. "resistance," obtained by dividing the d.c. arc voltage by the current, is included. This resistance, except for the longer lengths, is seen to be of the same order of magnitude as the r-f value and has the same trend. It is probable that the rapid decrease in R at short lengths (less than 0.25 cm) is due to the constriction that occurs in the column near electrodes and to the presence of some metal vapor in that

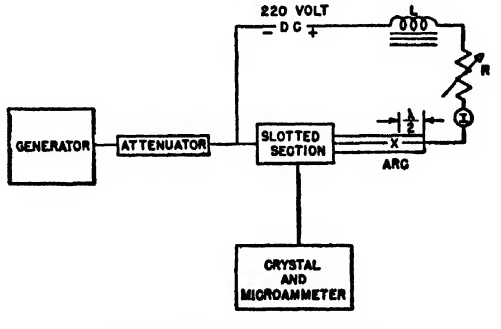


FIG. 2. Diagram of apparatus.

* Now at Electronics Laboratory, Electronics Department, General Electric Company, Syracuse, New York.
† Now at the Knolls Atomic Power Laboratory, General Electric Company, Schenectady, New York.

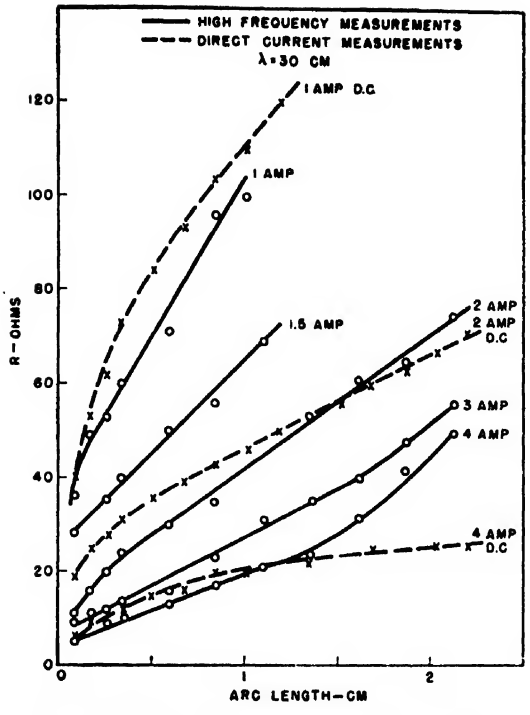


FIG. 3. Resistance characteristics of d.c. arcs in air.

legion. The upward trend of the r-f resistance of the ronger gap lengths may be due to the presence of hot ionized gas flowing along the half-wave section of the upper electrode. This would introduce losses in that section, increasing the measured value. The fact that this trend is most pronounced at the higher current is in agreement with this concept.

The arc consists of three zones of markedly different characteristics. At the cathode is a positive ion space-charge zone having a d.c. voltage drop of the order of the ionization potential of the gas. At the anode is an electron space charge zone that may have a voltage drop as high as the ionization potential of the gas. Both of these space-charge zones are of the order of a few mean free paths ($\sim 10^{-4}-10^{-5}$ cm) thick and the current density[1] at each is of the order of 50,000 amp./cm². Between these small zones is the column, which is of essentially uniform current density, except for constrictions near the ends where space charge zones develop. The current density for the low current arc column is approximately uniform along the column and is of the order of 10 amp./cm². The column gradient is uniform and relatively low. It is evident that a "resistance" obtained by dividing the total arc drop by the current might be considerably different from that measured by the r-f. The entire r-f resistance might be assumed to reside in the column. In this case, the d.c. column resistance is obtained by subtracting a reasonable maximum value of anode plus cathode voltage drops (15+15) from the total voltage and dividing by

[1] J. D. Cobine and C. J. Gallagher, Phys. Rev. 74, 1524 (1948).

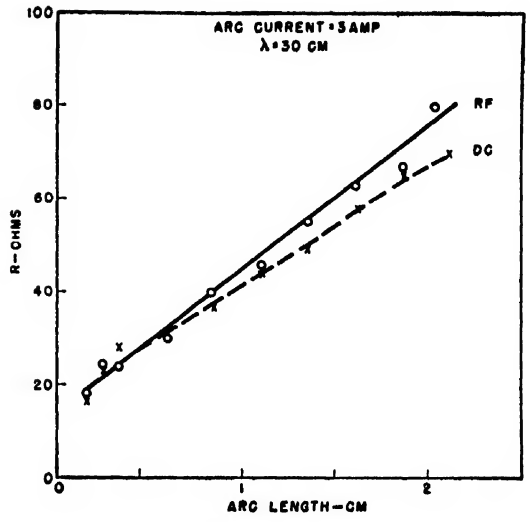


FIG. 4. Resistance characteristics of d.c. arcs in helium.

the current. The resulting resistance curves are not as good in absolute value as those shown in Fig. 3, except for 2 amperes where r-f and d.c. values are very nearly the same throughout most of the range in length studies. However, the slopes of the d.c. resistance *vs.* length curves thus obtained were nearly the same as for the high frequency curves over most of the range. The greatest error here may be in assuming the anode drop to have its maximum value; actually, it may have been only a few volts. Few really reliable values of anode drop have been measured for the high pressure arc.

The r-f resistances of arcs in helium and argon are shown in Figs. 4 and 5. The helium characteristic does not curve upwards at the longer gaps as does that for

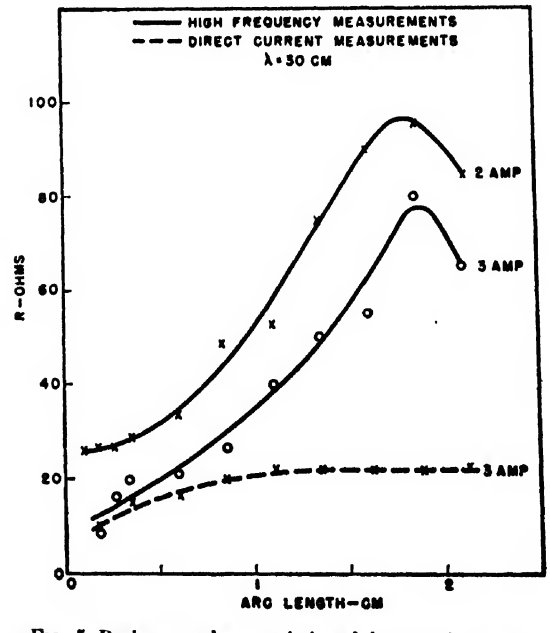


FIG. 5. Resistance characteristics of d.c. arcs in argon.

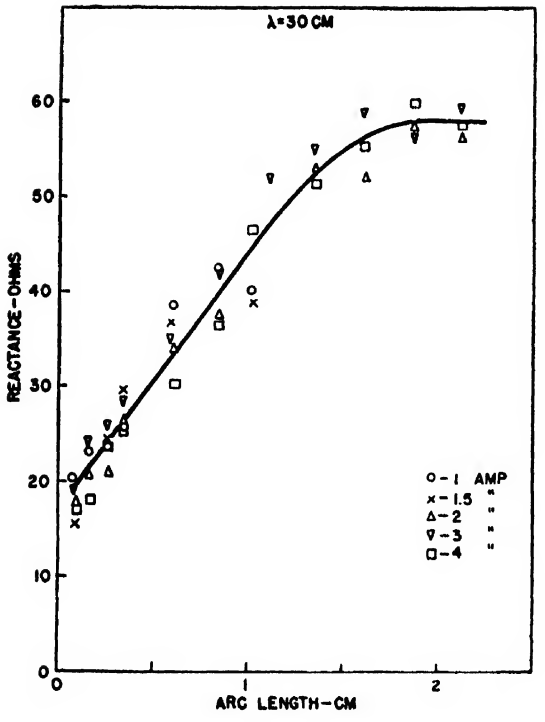


FIG. 6. High frequency reactance of d.c. arcs in air.

air. This might be the result of a lower temperature column of different size and more rapidly cooled residual gas than was true for air. The argon characteristic exhibits a definite maximum that was present in all the runs taken. There is no obvious explanation for this nor for the marked departure of the d.c. resistance from the r-f at the longer gap lengths.

The effects of arc current and gap length on the reactance of the d.c. arc in air are shown in Fig. 6. The reactance is capacitive. It is evident that arc current has very little effect on the reactance measured. Probably the principal effect in the straight line portion is due to the increase in column diameter as current is increased. There seems to be a systematic trend at the longer gap length for each current where the data indicate a departure from linearity; in fact, there may be a maximum near the end for each current. This departure occurs at increasingly greater gap lengths as the current is increased. There is no obvious explanation for this at present. The reactances for helium, air, and argon are compared in Fig. 7 for a 3-ampere arc. The general characteristics are the same for each gas. A definite maximum appears in the curves for helium and argon.

COMPARISON WITH THEORY

The fact that an atmospheric pressure arc in the low current range studied exhibits a capacitive reactance at high frequencies may be shown to be in agreement with the analysis of discharge admittance made by Everhart and Brown.[2] According to their analysis, a high conductivity discharge will have a capacitive reactance when the ratio

$$\gamma=\omega/\nu_c\leqq 0.25,$$

where ω is the angular frequency of the r-f and ν_c is the collision frequency of the electrons. The collision frequency is the ratio of the average velocity of the electrons to their mean free path. The average velocity may be taken as the average thermal velocity $\bar{c}$ of the electrons. If the arc temperature is assumed to be 6000°K, the value of $\bar{c}$ is 4.8×10^7 cm/sec. The probability of collision for electrons[3] of about this energy is 30 for N_2 and 20 for O_2. The gases in a high pressure arc column have a rather high degree of dissociation so that Brode's data are not strictly applicable. However, there seem to be no data for the probability of collision in atomic nitrogen or oxygen. Using the value for N_2 and correcting to 760 mm and 6000°K gives an electron mean free path $L_e=9.65\times10^{-4}$. Then

$$\nu_c=\bar{c}/L_e=4.8\times10^7/9.65\times10^{-4}=4.97\times10^{10}.$$

For an r-f frequency of 10^9 c.p.s. this gives

$$\gamma=2\pi10^9/4.97\times10^{10}=0.126.$$

Thus the conditions for a capacitive reactance are satisfied, provided the discharge is of high conductivity. The high frequency conductivity of the plasma can be calculated from the analysis of Margenau[4] as follows. The quantity $x_1(=m(\omega L_e)^2/2kT$, Reference 4, Eq. 20)

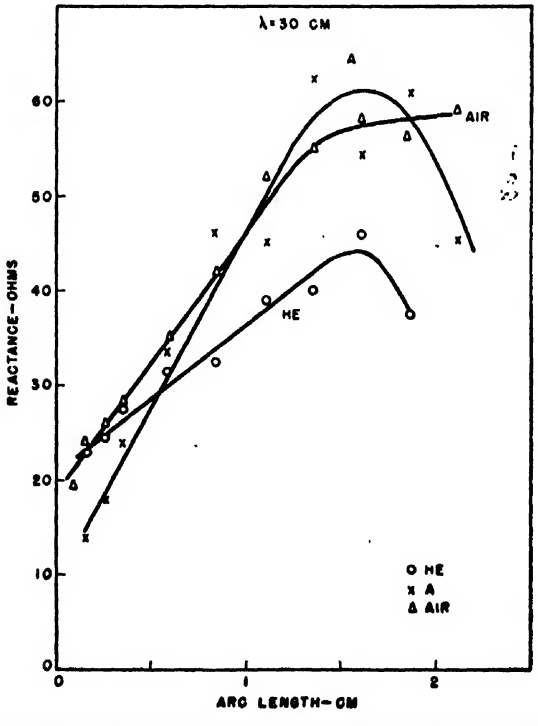


FIG. 7. High frequency reactance of 3-ampere d.c. arcs in helium, argon, and air.

[2] E. Everhart and S. C. Brown, Phys. Rev. 76, 839 (1949).

[3] R. B. Brode, Rev. Mod. Phys. 5, 257 (1933).

[4] H. Margenau, Phys. Rev. 69, 508 (1946).

has the value 2.02×10^{-3} when the above value of L_e is used and the electrons are assumed in thermal equilibrium at 6000°K. Thus the limiting values of $K_2=1$ and $K_1=\frac{1}{2}\pi^{\frac{1}{2}}$ may be used. The complex conductivity then becomes

$$\sigma=(ne^2/\omega m)[(4x_1^{\frac{1}{2}}/3\pi^{\frac{1}{2}})-j(2x_1/3)].$$

This may be compared with the free space susceptance by dividing by $\omega\epsilon_0$. Substituting the numerical values gives

$$\sigma/\omega\epsilon_0=8.62\times10^{-12}n-j1.09\times10^{-12}n.$$

If n is assumed to be of the order of 10^{14} for the high pressure arc[5]

$$\sigma/\omega\epsilon_0=8.62\times10^{2}-j1.09\times10^{2}.$$

Therefore the value of the Everhart and Brown constant η ($=|\sigma_{\omega 0}|/\omega\epsilon_0$) is of the order of 1000 and the discharge lies in the capacitive zone of their Fig. 2. Thus, the arc appears to be an example of a discharge having a positive susceptance, although the imaginary part of the complex conductivity is negative. Similar conditions were found at lower pressures by Everhart and Brown.[2] Their theory seems to agree with the observed results even though it was derived for a sinusoidal distribution of electrons along the arc, whereas in the arc the distribution is essentially uniform. A considerable discrepancy exists between the observed and theoretical values of conductivity which is not understood.

[5] C. G. Suits, Physics 6, 190, 315 (1935).

ACKNOWLEDGMENTS

Mr. R. H. Johnston took much of the data included in this paper. The coaxial line approach was suggested by Dr. M. D. Fiske of this Laboratory. Helpful suggestions were made by Professor Sanborn C. Brown of M.I.T. and by R. A. Dehn and Dr. C. J. Gallagher of this Laboratory.

A Relationship between Resistance and Temperature of Thermistors

G. Bosson, F. Gutmann, and L. M. Simmons
The New South Wales University of Technology, Sydney, Australia

(Received May 22, 1950)

The equation $\log R=A+B/(T+\theta)$ for the resistance R of a thermistor at T°K is proposed. Least square analyses of the most precise resistance-temperature data available for three different thermistor materials show that the equation is a considerable improvement over its predecessors; the standard relative errors of fit are 0.4 percent, 0.17 percent, and 0.91 percent.

THE $1/T$ law for thermistors

$$\log R=A+B/T \qquad (1)$$

derived from consideration of the free energy of an electron gas,[1] is known to be only approximate, since plots of $\log R$ *versus* $1/T$ yield bowed lines. By assuming that the slopes of these plots increase linearly with temperature, Becker, Green, and Pearson[2] proposed the following improved relationship, to which we shall refer as the B.G.P. law:

$$\log R=\log A-C\log T+B/2.303T. \qquad (2)$$

We have carried out a least square analysis of this law, using data kindly supplied by Bell Telephone Laboratories relating to the Western Electric Thermistor Material No. 1, the precision being: temperatures ±0.02°K, resistances ±0.02 percent. The results, shown in Table I, indicate clearly that considerable deviations occur between the resistances calculated from the B.G.P. law (R_b) and those observed (R_0).

TABLE I. Comparison between the goodness of fit of the B.G.P. and proposed laws according to data for Western Electric Thermistor No. 4756-12 (No. 1 Material).

		B.G.P. law		Proposed law	
Absolute temperature, °K	Observed resistance r_0 megohms	Residuals r_b-r_0 megohms	Percentage discrepancy $100(r_b-r_0)/r_0$	Residuals r_e-r_0 megohms	Percentage discrepancy $100(r_e-r_0)/r_0$
204.71	10.5587	+1.755	+16.6	+0.0646	+0.61
209.17	7.3683	+0.9516	+12.9	+0.0778	+1.06
216.95	4.1301	+0.2344	+ 5.68	−0.0076	−0.18
224.16	2.4654	+0.0330	+ 1.34	−0.0089	−0.36
232.54	1.3988	−0.0341	− 2.44	−0.0072	−0.51
240.05	0.86497	−0.04223	− 4.88	−0.00485	−0.56
247.74	0.54051	−0.03459	− 6.40	−0.00151	−0.28
257.18	0.31504	−0.02521	− 8.00	−0.00140	−0.44
266.48	0.19043	−0.01640	− 8.61	−0.00051	−0.27
276.34	0.11521	−0.00999	− 8.67	−0.00006	−0.05
286.27	0.071585	−0.005931	− 8.29	+0.000087	+0.12
296.35	0.045383	−0.003365	− 7.41	+0.000167	+0.37
308.54	0.027163	−0.001695	− 6.24	+0.000083	+0.31
320.24	0.017104	−0.000778	− 4.55	+0.000075	+0.44
331.57	0.011289	−0.000350	− 3.10	+0.000004	+0.04
344.07	0.0072974	−0.0000463	− 0.63	+0.0000142	+0.19
357.16	0.0047611	+0.0001019	+ 2.14	+0.0000120	+0.25
369.36	0.0032825	+0.0001552	+ 4.73	+0.0000034	+0.10
382.29	0.0022638	+0.0001750	+ 7.73	0.0000000	0.00
394.74	0.0016157	+0.0001745	+10.8	−0.0000015	−0.09
407.17	0.0011776	+0.0001662	+14.1	−0.0000044	−0.37

Standard relative error of fit:

$$\left(\frac{\Sigma(r_b-r_0/r_0)^2}{21}\right)^{\frac{1}{2}}=0.0804;\quad \left(\frac{\Sigma(r_e-r_0/r_0)^2}{21}\right)^{\frac{1}{2}}=0.00395.$$

r_b = resistance calculated from the B.G.P. law.
r_e = resistance calculated from the proposed law.

[1] N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, London, 1948), second edition, p. 157.
[2] Becker, Green, and Pearson, Elec. Eng. Trans. 65, 713 (1946).

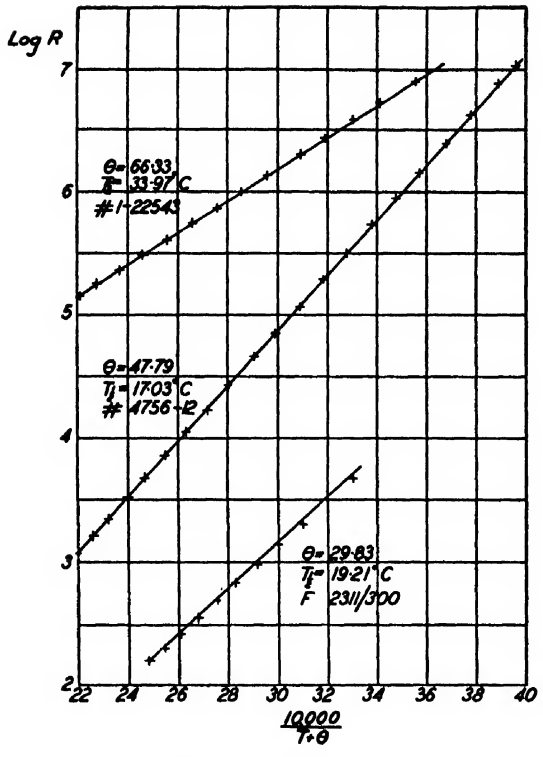


FIG. 1. Plots of log$R$ *versus* $10^4/(T+\theta)$ for three different thermistor materials.

The following empirical law is now proposed:

$$\log R = A + \frac{B}{T+\theta}. \qquad (3)$$

It indicates a linear relationship between log$R$ and $1/(T+\theta)$, and this is shown in Fig. 1 to hold for the three thermistor materials investigated, viz. Western Electric Material No. 1, Western Electric Zinc Ferrite No. 1-22543, and Standard Telephones and Cables, Ltd. Thermistor Type F2311/300. The values for θ were obtained by least square analyses of data supplied by Bell Telephone Laboratories, N. J., and Standard Telephones and Cables Ltd., London. These results are shown in Table II, which further demonstrates that Eq. (3) permits the resistance of a thermistor to be predicted from its temperature sufficiently closely for all but the most precise work.

TABLE II. The goodness of fit of the proposed law with data for three different thermistor materials.

°K	Resistance observed for thermistor No. 4756-12 megohms	No. 1-22543 megohms	F2311/300 ohms	Residuals in ohms	Percentage discrepancy
204.71	10.5587			+64600	+0.61
209.17	7.3683			+77800	+1.06
215.32		7.8491		+ 1700	+0.02
216.95	4.1301			− 7600	−0.18
224.16	2.4654			− 8900	−0.36
226.80		5.1994		−23000	−0.43
232.54	1.3988			− 7200	−0.51
236.29		3.7655		− 7800	−0.21
240.05	0.86497			− 4850	−0.56
247.21		2.6609		+ 1200	+0.05
247.74	0.54051			− 1510	−0.28
256.80		2.0044		+ 600	+0.03
257.18	0.31505			− 1400	+0.44
266.48	0.19043			− 510	−0.27
271.16		1.3499		+ 1700	+0.13
273.16			4733	+ 38	+0.80
276.34	0.11521			− 60	−0.05
284.04		0.97377		+ 1600	+0.17
286.27	0.071585			+ 87	+0.12
293.16			2003	− 5	−0.25
296.01		0.73361		+ 1800	+0.25
296.35	0.045383			+ 167	+0.37
303.16			1369	+ 5	+0.37
308.54	0.027163			+ 83	+0.31
309.47		0.54624		+ 740	+0.14
313.16			946	− 2	−0.21
320.24	0.017104			+ 75	+0.44
323.16			672	0	0.00
324.72		0.40050		+ 370	+0.09
331.57	0.011289			+ 4	+0.04
333.16			485.3	− 0.8	−0.16
339.85		0.30144		− 60	−0.02
343.16			354.7	− 2.9	−0.82
344.07	0.0072974			+ 14.2	+0.19
353.16			267.3	0	0.00
355.14		0.23082		− 180	−0.08
357.16	0.0047611			+ 12	+0.25
363.16			201.5	− 1.3	−0.65
369.36	0.0032825			+ 3.4	+0.10
371.98		0.17580		− 2600	−0.15
373.16			160.0	+ 4.0	+2.50
382.29	0.0022638			0	0.00
386.29		0.14156		− 120	−0.08
394.74	0.0016157			− 1.5	−0.09
407.17	0.0011776			− 4.4	−0.37
Standard relative error of fit:	0.00395	0.00170	0.00908		

TABLE III. Thermistor constants.

Thermistor	A	B°K	θ°K	$T_{\frac{1}{2}}$
S.T.C. F2311/300	−2.28697	1805.41	29.83	19.21
Zinc Ferrite 1-22543	+2.27698	1300.64	66.33	33.97
No. 1 Material 4756-12	−1.86554	2245.18	47.79	17.03

Table III shows the values of the constants A, B and θ for the three thermistors discussed, together with their half-temperatures at $T_{\frac{1}{2}}$ 300°K.

A High Intensity Short Duration Spark Light Source*,**

J. A. Fitzpatrick, J. C. Hubbard, and W. J. Thaler
Department of Physics, Catholic University of America, Washington, D. C.
(Received May 25, 1950)

The spark described here provides a high intensity light source, whose effective duration approaches 10^{-7} second. The light intensity is nine hundred times greater than that obtained using RG8U cable cut for 10^{-7} second duration discharge. The cable described here is coaxial with silver conductors fired on the dielectric, a compound of barium titanate. The gain in light intensity is apparently due to both the higher capacity of the barium titanate cable and its very low characteristic impedance.

A COAXIAL cable compounded of barium titanate provides a spark light of both high intensity and short duration. This is accomplished by applying the method employed by Beams *et al.*[1] to a cable of low characteristic impedance. Since the velocity of propagation of the discharge wave is low, a short length of cable, 16.5 cm is used. The cable is cylindrical, the outside conducting surface having a diameter of five centimeters and the inside a diameter of two and one-half centimeters, with the dielectric forming the body of the cable.

The discharge producing the light is confined to a small hole made with a number 80 drill, 0.013-inch diameter and 4-mm long drilled in a piece of unfired soapstone, Fig. 1. For some applications a triggering gap is fitted in series with the light producing gap to discharge the cable at a specific time.

The dielectric constant of the material forming the cable has been measured at frequencies up to 250 megacycles. A General Radio Impedance Bridge was used for the low frequency measurement at 1000 c.p.s. For the higher frequencies resort was had to using the cable as a shorted and open end transmission line, and observing the resonance frequencies indicated on a grid dip meter connected as a generator to the sending end of the line. For a short circuited line the impedance presented to the generator is zero for an infinitely short line, becoming an increasing inductive reactance as the line is lengthened until the impedance becomes infinitely great for a value of $2\pi l/\lambda=\pi/2$ when l, the length of the cable, equals one quarter wave-length in the dielectric. Since the length l of the cable was fixed at 16.5 cm, the frequency of the generator was varied between a range of 2 to 250 megacycles and the frequencies of resonance or maximum impedance observed as the patterns of standing waves on the line were established. Knowing the frequency and wave-length, the velocity of propagation was determined as well as the dielectric constant. The values shown in Table I were obtained for both an open and short circuited line.

The characteristic impedance of the line is a function of the dielectric constant. Since a cable will discharge in a square wave of current if terminated in its characteristic impedance, it is desirable to make such a termination. We see from Table I that it is impossible to assign a fixed value to the characteristic impedance which will satisfy all the frequencies present in the discharge. The terminal impedance corresponding to the lowest value of the dielectric constant is 1.5 ohms while the value for the highest is 1.2 ohms. The writers have inserted various resistors of between one and two ohms in series with the gap but were unable to observe any effects on shadowgraphs made of progressive ultrasonic waves such as shown in Figs. 2–4, even when terminating resistors were eliminated. It is believed that the resistance of the fittings of the spark head as well as some resistance presented by the spark gap are close enough to the characteristic impedance of the cable to allow the omission of a resistor in series with the gap.

The duration of light produced by the barium titanate cable has been compared with a section RG8U cable cut for 10^{-7} second discharge, approximately 9½-meters long. Both discharges were made in confined gaps. Figure 5 shows a cathode ray oscilloscope trace and a five megacycle time marking signal superimposed on the discharge. These traces were made with a photo-

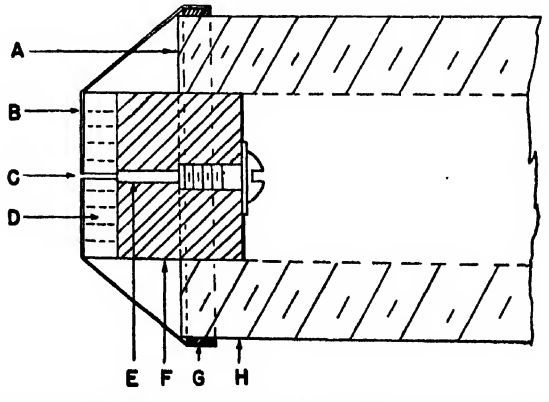


Fig. 1. Detail of light source. *A*, cylinder of barium titanate dielectric; *C*, gap space, hole of 0.013-inch diameter in unfired soapstone cylinder, *D*, between stainless steel electrode, *B*, and tungsten electrode, *E*, fitted to an adjusting screw; *F*, polystyrene support; *G*, clamp to silver outer conductor, *H*, of the cable.

* Work supported by ONR.
** Presented at A.P.S. meeting, Washington, D. C., April 28, 1950.
[1] Beams, Kuhlthau, Lapsley, McQueen, Snoddy, and Whitehead, Jr., J. Opt. Soc. Am. 37, 868 (1947).

TABLE I. Measured value of the dielectric constant of the barium titanate compound forming the coaxial cable.

Resonant frequency megacycles	λ(cm)	Velocity of propagation (cm/sec.) $\times 10^{-8}$	Dielectric constants
(a) Shorted line			
15.5	66	10.2	860
50	22	11.0	745
75	13.2	9.9	925
105	9.44	9.9	925
140	7.32	10.25	854
167	6.0	10.0	900
190	5.08	9.64	970
250	4.4	11.0	745
(b) Open line			
26.9	33	8.88	1140
76	11	8.36	1290
143	6.8	9.72	955
187	4.71	8.80	1162
240	3.67	8.80	1162
(c) 1000 c.p.s. bridge (G.R. type 650A)			1575

multiplier tube and synchronoscope combination.*** The following comparative data are obtained.

	RG8U Cable	$BaTiO_3$ Cable
Over-all light duration	7×10^{-7} sec.	12×10^{-7} sec.
Decay time, peak to $1/e$ peak	1.9×10^{-7} sec.	2×10^{-7} sec.
Rise time, zero to peak	1×10^{-7} sec.	1×10^{-7} sec.
Fifty percent of peak limits	1.5×10^{-7} sec.	1.8×10^{-7} sec.

The series of spark shadowgraphs shown in Figs. 2, 3, and 4 show the approach to a resolution limit as the frequency of the progressive ultrasonic waves is increased from three to nine megacycles. A quartz crystal having a fundamental frequency of one megacycle was driven at 3, 7, and 9 times its fundamental frequency to obtain these shadowgraphs. The method used was that devised by Hubbard, Zartman, and Larkin.[2]

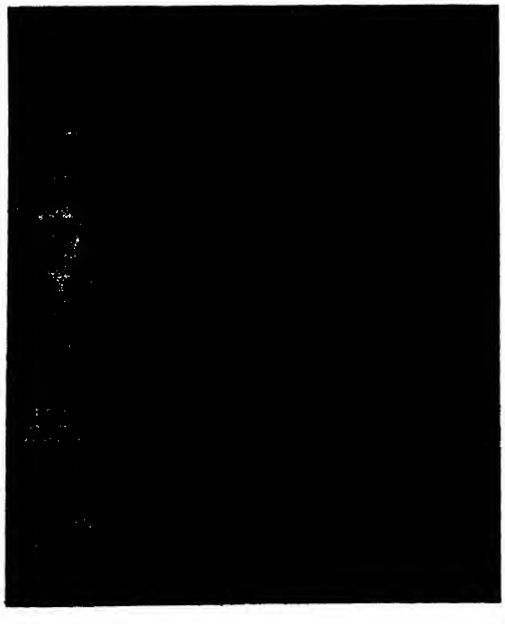


FIG. 2. Shadowgraph of 3.0 Mc progressive ultrasonic wave in water. The scale shown is in cm. Water temperature—25°C.

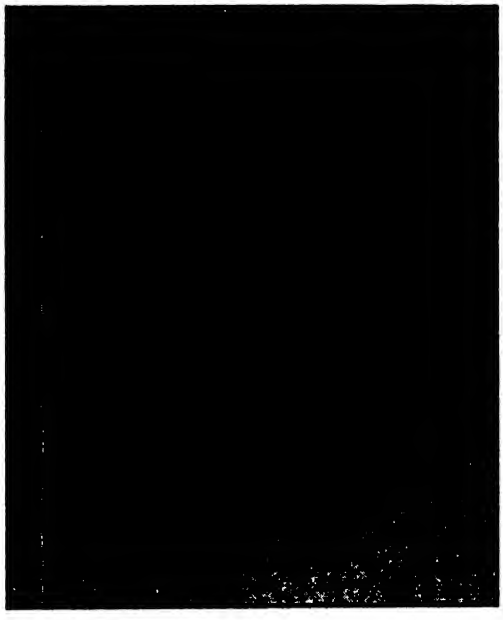


FIG. 3. Shadowgraph of 7.0 Mc progressive waves in water. Note that definition is not as good as that shown in Fig. 1. See also caption in Fig. 4.

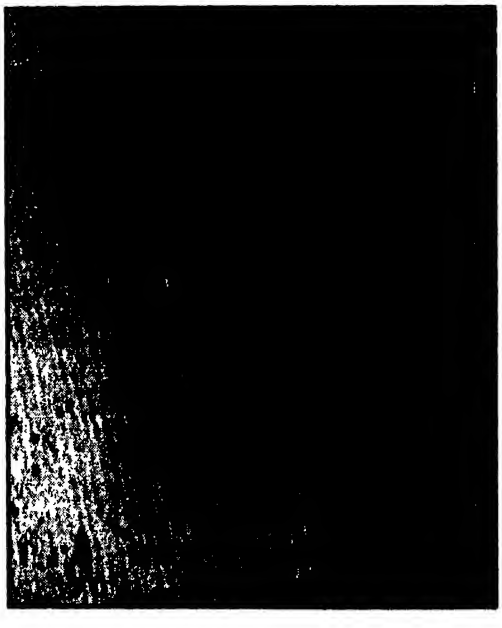


FIG. 4. Shadowgraph of 9 Mc progressive wave in water. The wave pattern here is not very distinct indicating that the acoustic wave has moved an appreciable part of a wave-length during the period of illumination. A diminution of the sound intensity is seen in the short distance covered by this picture. The marginal spacing is in mm.

*** These traces were made through the courtesy of Mr. Wm. T. Whelan of the Naval Ordnance Laboratory.

[2] Hubbard, Zartman, and Larkin, J. Opt. Soc. Am. 37, 832 (1947).

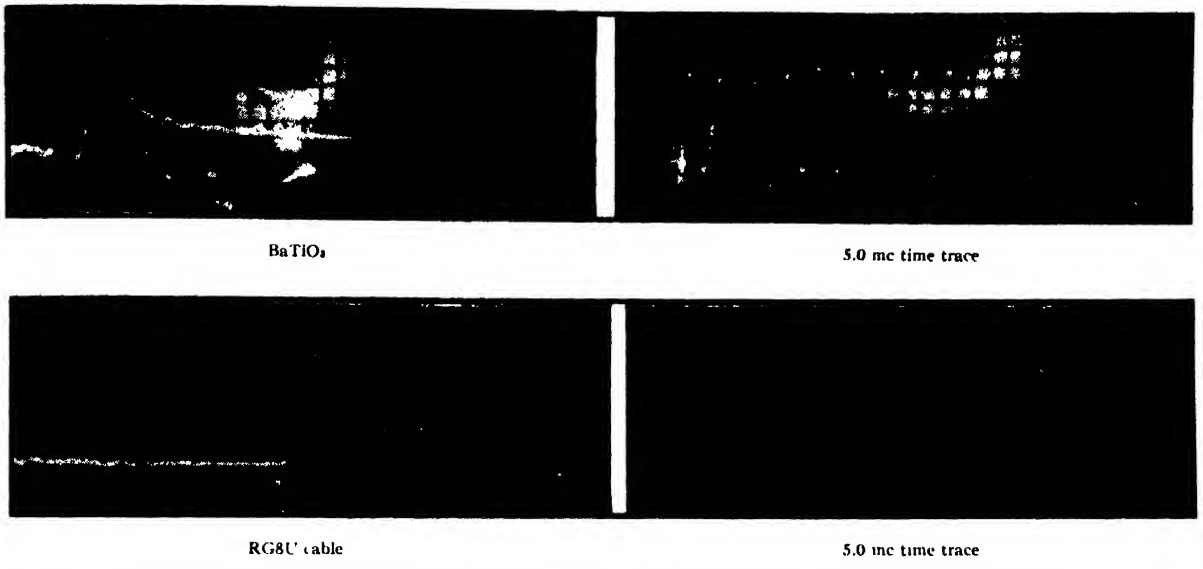


FIG. 5. Photograph of cathode-ray oscilloscope traces of light from spark sources as a function of time. The time scale is given by the sinusoidal traces on the right of frequency 5 Mc/sec.

Comparative light intensities have been made by discharging both the RG8U cable and the $BaTiO_3$ cable through the same spark electrodes and allowing the divergent light to fall on the sensitive plate of a photomultiplier tube the output of which was coupled to a ballistic galvanometer. To obtain the same deflection from the galvanometer as was obtained from discharging the RG8U cable at a fixed distance it was necessary to move the barium titanate light source to 30 times the distance from the photo-cell, thus indicating a light intensity nine hundred times as great as that produced by the RG8U cable. It is possible to obtain good spark photographs on lantern slide plates four feet away from the barium titanate divergent light source. Apparently the nine-hundred fold gain in light intensity is due to the elimination of the loss inherent in the fifty ohm terminating resistor associated with the RG8U cable since at the most, the electrical capacity of the barium titanate cable is only twenty times that of the RG8U cable used here.

The high light intensity and short time of the spark together with the convenient size of the cable make it particularly useful for studying high speed phenomena such as turbulence and ultrasonic wave-systems where photographic methods are used. We have observed while employing this device that the light intensity and spark duration are functions of many parameters principal of which are, gap spacing, current density in the gap, which depends on the initial charge on the line (the results shown here have been made, with the cable charged to 10,000 volts), and type of confining gap; a confined spark produces higher intensity but longer duration, an open gap less light intensity but shorter duration.

The authors wish to acknowledge with thanks the advice of Mr. Bert H. Marks, Product Engineer for Centralab Incorporated, which fabricated the cable.

Emissivity Changes of Thoria Cathodes*

O. A. Weinreich
Bartol Research Foundation of The Franklin Institute, Swarthmore, Pennsylvania
(Received June 1, 1950)

Spectral and total emissivity measurements of thoria-coated tungsten filaments show that irreversible and reversible changes of the emissivity occur. ϵ_λ ($\lambda=0.65\mu$) varies from 0.2 to 0.7 and ϵ_t from 0.2 to 0.5.

Certain variations of the thermal radiation can be related to changes of thermionic emission.

Passage of current through the coating affects emissivity and thermionic emission. This is shown by measuring infra-red radiation and electron current simultaneously. The existing relationship between emissivity and thermionic activation is discussed.

1. INTRODUCTION

THE numerous published values[1–7] of spectral emissivities for thoria vary over a wide range. (If thin coatings on metals are investigated, one should keep in mind that part of the radiation comes from the base metal.) The extinction coefficient at $\lambda=0.665\mu$ for thoria sleeves has been found to be 26 cm^{-1}.[8] The results of our investigations show that both reversible and irreversible changes in spectral and total emissivities can be effected by various treatments of the thoria.

Spectral and total emissivity measurements of thoria-coated tungsten cathodes are of interest in two respects:

(1) Total emissivity changes of thoria cathodes which occur during processing and life affect temperature and heater power; emissivity changes in the red affect the pyrometer corrections.

(2) Emissivity changes, insofar as they involve impurity levels, may yield information on the mechanism of thermionic emission.

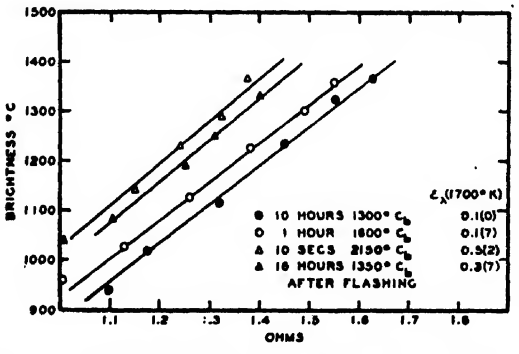


FIG. 1. Spectral emissivity ($\lambda=0.65\mu$) of a thoria-coated tungsten filament (coating thickness 50μ) after various heat treatments. True temperature is measured by filament resistance. Calibration curve not shown.

2. EMISSIVITY CHANGES OF THORIA-COATED W FILAMENTS

In an earlier paper[9] it was shown that a small spot of thoria on a tungsten filament appeared brighter than the tungsten, when previously flashed to ~2800°K. The brightness of the spot relative to the brightness of the tungsten decreased at lower temperatures with time. In the present investigation the emissivity changes of thoria-coated cathodes were investigated further by using thoria-coated tungsten filaments with potential leads. The filament resistance can be taken as a measure of the true temperature for the thoria surface if we assume that:

(1) The temperature drop in coatings $\lesssim 50\mu$ is negligible, which appears to be justified on account of the relatively high thermal conductivity of thoria, 0.01 cal. per cm deg.[10]

(2) The change of the tungsten filament resistance due to chemical reaction between tungsten and thoria is negligible. This has been shown to be true by evaporating completely a 10μ thoria coating from a tungsten filament. Before coating and after evaporation the resistance was the same within the experimental errors.

Tungsten filaments of 10 and 12 mils diameter and about 25 cm length were used. The potential leads of 2 mils diameter were spot-welded about 3 cm above the

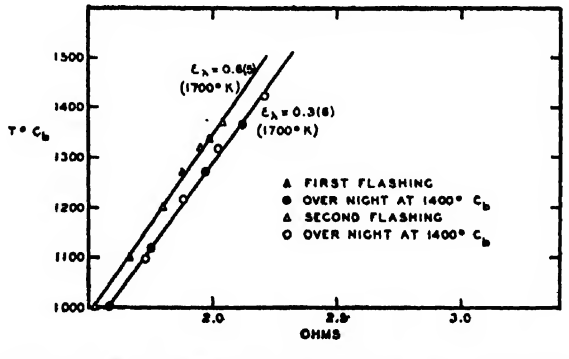


FIG. 2. Reversible spectral emissivity changes ($\lambda=0.65\mu$) of a thoria-coated tungsten cathode. Flashing to about 2300°C$_b$ increases the emissivity to $\epsilon_\lambda=0.7$ and subsequent heating at 1400°C$_b$ causes the emissivity to drop to $\epsilon_\lambda=0.4$.

* Part of the results described here have been presented at the New York Meeting of the Physical Society (Division of Electron Physics), January, 1950.

[1] Coblentz, Bur. Stand. Bull. 9, 283 (1913).
[2] Burgess and Waltemberg, Bur. Stand. Bull. 11, 591 (1915).
[3] W. E. Forsythe, Phys. Rev. 25, 252 (1925).
[4] G. Liebmann, Zeits. f. Physik 63, 404 (1930).
[5] D. A. Wright, Nature 160, 129 (1947).
[6] E. T. Hanley, J. App. Phys. 19, 583 (1948).
[7] S. Mesnard, Comptes Rendus 230, 1582 (1950).
[8] W. E. Danforth, private communication.
[9] O. Weinreich, Rev. Gen. d'El. 54, 243 (1945).
[10] W. E. Danforth, private communication.

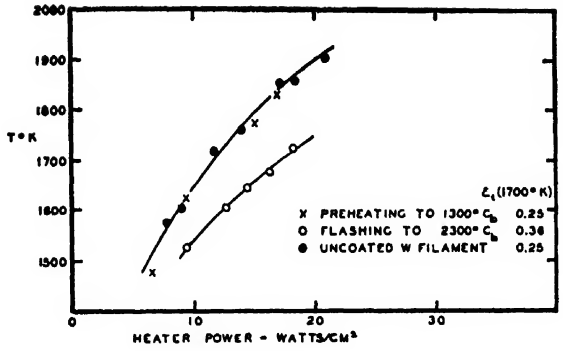


FIG. 3. Total emissivity of a thoria-coated tungsten filament. Coating thickness 10μ. Flashing to 2300°C$_b$ increases the total emissivity from $\epsilon_t=0.25$ to $\epsilon_t=0.36$ (at 1700°K).

filament terminals. The resistance was measured by the current through the filament and the voltage drop across the potential leads. The general procedure was the following: Filaments with potential leads were mounted in a tube which was pumped and baked in the usual manner, and the resistance-temperature characteristic of the uncoated tungsten filament was measured using pyrometer readings and corrections for the spectral emissivity of tungsten. The tube was then taken off the pumping system and opened, and a thoria coating was deposited on the filament. Commercial thoria powder supplied by Lindsay Light and Chemical Company was suspended in butyl alcohol, and the particles charged by addition of a small amount of thorium nitrate, as described by E. T. Hanley.[11] The tube was then sealed and put back on the pumping system. The resistance of the coated W filament was measured with the same instruments as used for the resistance-temperature plot of the pure tungsten filament.

Figure 1 shows the change of the apparent spectral emissivity of a thoria coating of 50μ thickness, when heated in steps to higher temperatures. The brightness temperature is plotted against the filament resistance. We see that heat treatment of the filament at increasing temperatures increases the spectral emissivity; i.e., for a given filament resistance the brightness temperature increases. (It should be noted that the initial emissivity increase may be partly an apparent increase due to a sintering effect, prior to which a temperature difference between filament and coating may exist.) The emissivity increase after flashing to ~2800°K is not stable but decreases with time if the filament is allowed to remain at lower temperatures, e.g., 1350°C$_b$.

The reversible change of the spectral emissivity of another filament is shown in Fig. 2. The emissivity increases, after flashing, to $\epsilon_\lambda=0.7$ and decreases with time, when held at 1400°C, to $\epsilon_\lambda=0.4$. This process can be repeated until finally the coating thickness decreases by evaporation during the flashing.

Changes of the total emissivity which occur with the same heat treatments are shown in Fig. 3. The heater

[11] E. T. Hanley, J. App. Phys. 19, 583 (1948).

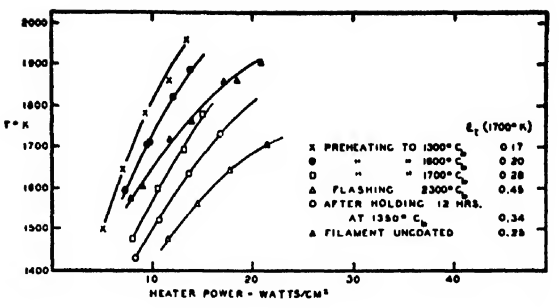


FIG. 4. Total emissivity of a thoria-coated tungsten filament (coating thickness 50μ) after various heat treatments.

power in watts/cm^2 is plotted against the temperature as measured by the filament resistance. The coating thickness is 10μ. The initial total emissivity at 1300°C$_b$ is very near to that of the pure tungsten, $\epsilon_t=0.25$. Again, as seen for the emissivity in the red, the total emissivity also increases, $\epsilon_t=0.36$, as a result of the flashing.

Figure 4 shows the same measurements for a 50μ coating. An initial heat treatment not exceeding 1300°C$_b$ gives $\epsilon_t=0.17$, which value increases after flashing to $\epsilon_t=0.45$. As the temperature is allowed to remain at 1350°C$_b$ the emissivity drops to $\epsilon_t=0.34$.

3. EMISSIVITY CHANGES AND ACTIVATION CHANGES OBSERVED WITH DIODES

As was shown in an earlier paper,[9] flashing to 2800°K increases electron emission and emissivity. Subsequent current drawing accelerates the deactivation, which normally occurs slowly with time after flashing if the cathode is maintained at a lower temperature. Here we

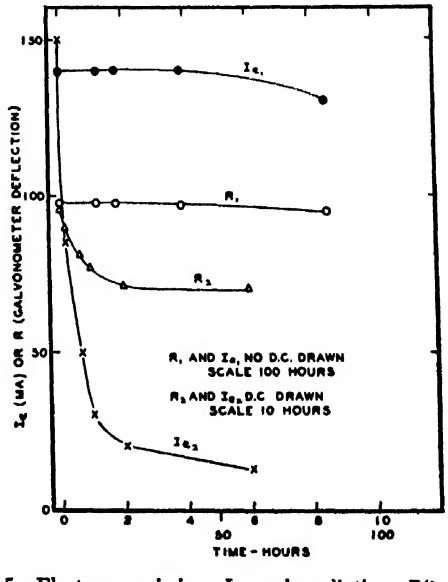


FIG. 5. Electron emission Ie and radiation $R(\lambda=1\mu)$ at $T=1175$°C of a thoria-coated tungsten filament as a function of time after flashing to 2300°C$_b$. Curves Ie_2 and R_2 show that the rates of decrease of Ie and R are highly increased if electron emission is drawn.

observe further the emissivity changes of a thoria cathode by measuring the intensity of radiation at an arbitrary wave-length in the infra-red with a thermopile, if the true temperature is kept constant by keeping the filament resistance at a constant value. Several diodes, the cathodes of which had potential leads, were made. The cylindrical anodes were made of tungsten mesh. In this way the decrease of the spectral emissivity with time, together with the thermionic deactivation after the flashing could be measured. The following

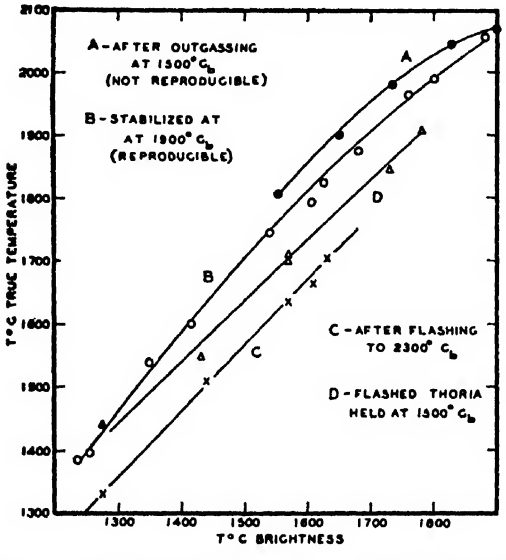


FIG. 6. Irreversible (curves *A* and *B*) and reversible (curves *C* and *D*) changes of the spectral emissivity ($\lambda=0.65\mu$) of a thoria-coated tungsten cathode after various heat treatments. The true temperature is measured by a Mo—W thermocouple.

observations were made at lower temperatures after flashing the cathode to ~2800°K:

(a) If no electron emission is drawn, both intensity of radiation and thermionic activity decrease slowly with time. The rate of this decrease depends on the temperature. For example, at 1400°C radiation and emission decreased about 10 percent after 100 minutes, while at 1150°C, no appreciable decrease was seen after 85 hours.

(b) The drawing of electron currents accelerates drastically the rate of deactivation and of emissivity decrease. For example, at 1400°C the electron emission decreased 50 percent and the radiation by 33 percent after ten minutes.

(c) The accelerated decrease of thermionic emission and emissivity due to the drawing of current could still be seen if, after flashing, the cathode was first held "quiescent" at 1400°C for 16 hours and then the anode voltage applied. The described phenomena were qualitatively reproducible and observed with several diodes and at different wave-lengths. Figure 5 shows the results for 1175°C at $\lambda=1\mu$.

4. EMISSIVITY CHANGES AS OBSERVED WITH A THERMOCOUPLE

Further verification of the emissivity changes of thoria was made using a W—Mo thermocouple,[12] spot-welded to a 20 mil tungsten filament to measure the true temperature. The known spectral emissivity of the uncoated tungsten filament was used to calibrate the e.m.f. of the thermocouple against true temperatures. A plot of the true temperature against the e.m.f. was made over a range from 1200° to 2100°C. The tube was then opened, the tungsten filament coated with thoria by cataphoresis (about 20μ thick), sealed, and put again on the pumping system. In Fig. 6 the brightness temperatures of the thoria coating are plotted against the true temperatures as measured by the thermocouple. After outgassing at 1500°C$_b$ the temperature of the filament was raised in steps to 1900°C$_b$ (curve *A*). Curve *B*, obtained by decreasing the temperature from 1900°C to 1200°C does not coincide with curve *A* and is the reproducible curve. In contrast, curve *A* is not reproducible. This behavior is believed to be the result of a certain sintering action, which occurs along curve *A* and is completed at 1900°C. (The very low initial value of the spectral emissivity of thoria coatings ($\epsilon_\lambda=0.09$) found recently by S. Mesnard[7] may be an apparent one for the same reason.) Following flashing to 2300°C$_b$ temperature measurements gave curve *C*. After several hours at 1500°C$_b$, curve *D* is obtained. The spectral emissivity values ($\lambda=0.65\mu$) for the different states, taken from Fig. 6 for $T=1700$°C are:

Before flashing: $\epsilon_\lambda=0.3$ (curve *B*).
After flashing: $\epsilon_\lambda=0.7$ (curve *C*).
After flashing and 10 hours at 1500°C$_b$: $\epsilon_\lambda=0.4$ (Curve *D*).

The tube was now sealed off, and after a new flashing to 2300°C$_b$ the radiation of the thoria-coated filament at $\lambda=1\mu$, as obtained from the galvanometer deflections of the thermopile of the spectrograph, was plotted against time for a temperature of 1520°C, this temperature being maintained constant by holding the e.m.f. of the thermocouple on the filament at a constant value. Again, the emissivity decreased with time as was expected from the previous measurements when the true temperature was measured by the resistivity.

5. CONCLUSION

It is premature to attempt to draw theoretical conclusions from the results presented here. However, apart from the bearing on practical problems of heater power and pyrometer corrections, it seems that the results show the existence of a relationship between intensity of thermal radiation and thermionic activation.

Undoubtedly chemical reactions occur between tungsten and thoria when flashed to 2800°K. Reaction products may be metallic thorium and/or some thorium

[12] F. H. Morgan and W. E. Danforth, J. App. Phys. 21, 112 (1950).

suboxide. W. L. B. Burgers and T. M. Liempt[13] have made x-ray analyses of thoriated tungsten filaments. They found that when these tungsten filaments are heated to 2500°K, thoria is partly reduced to metallic thorium. Besides this, their results show some evidence for incomplete reduction and formation of an oxide with an oxygen content between Th and ThO_2. The authors found a similar reaction product when heating superficially oxidized thorium rods. The only impurity considered so far for thoria cathodes is free thorium, and many activation phenomena fit into this picture.[14] On account of Burgers and Liempt's work, one should also consider the possibility of suboxide impurities.

Some doubt may be held as to whether or not a density of impurity centers, sufficient to account for conduction and thermionic emission, is also sufficient to provide observable changes of optical emission. It is possible that our reversible emissivity changes might be linked to thermionic activation in an indirect way only, i.e., impurity material produced in such quantity that colloidal agglomeration occurs. Present views concerning thoria[15, 16] indicate a density of impurity centers of the order of 10^{18} cm^{-3}. This figure, although its value is subject to further investigation, is of the same order as the density of color centers of the alkali halides in which measurable optical effects are found. It therefore seems possible that knowledge regarding impurity levels in thoria might be gained from infra-red emission studies, and further work is in progress in this connection.

ACKNOWLEDGMENT

The writer wants to thank Mr. F. H. Morgan for his aid in assembling and calibrating the W—Mo thermocouples and for his assistance in carrying out the experiments described in Sec. 4.

[13] W. L. B. Burgers and T. M. Liempt, Zeits. f. anorg allge. Chemie **193**, 144 (1930).
[14] H. Y. Fan, J. App. Physics **20**, 682 (1949).
[15] D. A. Wright, Proc. Phys. Soc. **B62**, 188 (1949).
[16] W. E. Danforth and F. H. Morgan, Phys. Rev. **79**, 142 (1950).

Techniques for Measuring the Dynamic Characteristics of a Low Pressure Discharge

BENTLEY T. BARNES AND STEPHEN EROS
General Electric Company, Lamp Development Laboratory, Nela Park, Cleveland, Ohio

(Received June 13, 1950)

The dynamic characteristics of a discharge can be determined by making square-wave changes of arc current. This is done by putting the discharge tube in the plate circuit of a group of pentodes connected in parallel and supplied with a square-wave signal superimposed on a d.c. bias.

Probes operated at constant current are very useful in recording the characteristics of a modulated discharge. For measuring electron temperature (T_e), two identical probes are centered about points having the same space potential. If the probes are operated at different fixed currents on the straight line part of the probe characteristic, the voltage between them is proportional to T_e. Longitudinal voltage gradient or radial potential differences may be recorded by use of identical probes operated at the same current.

A probe operated at a suitable current will follow roughly the variations in space potential of the adjacent region of the discharge. Another probe made sufficiently negative with respect to such a probe will collect only positive ions. A probe made 0–10 volts positive with respect to the constant-current one gives records from which electron concentrations can be computed. In this case, a bias connection without current flow is obtained by use of a cathode-follower with two stages of direct-coupled amplification.

The oscillograph amplifier used for dynamic probe records must be either a differential one or one with ungrounded power supply. In the latter case, the stray capacitance is many times larger; but fairly reliable records can be obtained if the arc current is 0.1 ampere or more, and if the time constant for the decay of the transient effects being measured is at least 10 microseconds. The oscilloscope amplifier and any preamplifier used for recording spectral intensities should be direct-coupled.

IF a discharge is operated on a.c. at a frequency of 1000 c.p.s., or more, the volt-ampere curve obtained on an oscilloscope screen is quite different from the curve obtained by plotting voltage readings taken on a d.c. discharge operated at various currents. This difference in behavior shows that the dynamic characteristics of a discharge are not the same as those indicated by a series of static tests. Since this is of considerable interest from both practical and theoretical viewpoints, we have developed methods for getting fairly complete data on the dynamics of a discharge. The latter is subjected to square-wave changes of arc current. The time between successive square-wave changes should, of course, be long enough to permit the discharge to approach steady-state conditions. This permits one to observe both the instantaneous effect of a change in current and the rate of adjustment toward the steady state. To make measurements under these conditions, one needs to modify the techniques developed by Langmuir[1,2] for measurements on discharges operated at a constant current. In this article we describe equipment and procedures for obtaining square-wave changes of arc current and for recording the instantaneous values of various properties of the discharge.

SQUARE-WAVE MODULATION

The circuit used for controlling arc current is shown in Fig. 1. We have used twelve 6L6 tubes, connected in

[1] I. Langmuir, G. E. Rev. **26**, 731 (1923).
[2] I. Langmuir and H. Mott-Smith, Jr., G. E. Rev. **27**, 449 (1924).

parallel except for a non-inductive resistor in series with each grid. These resistors prevent parasitic oscillations. They are connected to the cathode-follower output stage of a square-wave generator. To provide the proper bias for the 6L6's, the negative terminal of the voltage supply for the square-wave generator is put at −60 volts with respect to the negative lead of the voltage supply for the circuit shown in Fig. 1. Figure 2 shows how provision is made for controlling both the d.c. bias and the amplitude of the square-wave signal. This permits one to adjust the modulation and the average arc current independently.

With the circuit shown in Fig. 1, inductance in the leads of the voltage supply tends to round the corners of the square wave. Although the plate current of a pentode is insensitive to changes in plate voltage, it is affected to a much greater extent by changes in screen voltage. A better wave-shape would be obtained if a separate low-impedance supply were used for the screens of the 6L6's. This improvement may not be needed in a study of the dynamics of the flow of ions to the tube wall, but it is essential if one wishes to record the increase in ion concentration when the current is suddenly increased. In the latter case, a square wave with a rise time which does not exceed a few microseconds is required for even a rough separation of instantaneous effects from subsequent adjustments toward steady state conditions.

We have not tested the circuit change suggested in the preceding paragraph, partly because it did not occur to us until after a fairly complete set of data had been taken, and partly because our oscilloscope equipment could not resolve the details of changes which take place within a few microseconds. With a common plate-screen supply, 90 percent of a square-wave change in lamp current was made in about 80 microseconds. With the lamp current zero, the corresponding change in the grid voltage of the 6L6 tubes took less than one-third of this time.

PROBE TECHNIQUES

Sudden large changes in arc current are accompanied by proportionately larger instantaneous changes* in the

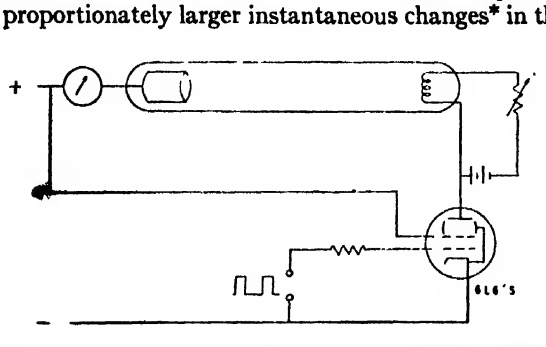


FIG. 1. Circuit diagram for discharge tube with square-wave modulation of arc current.

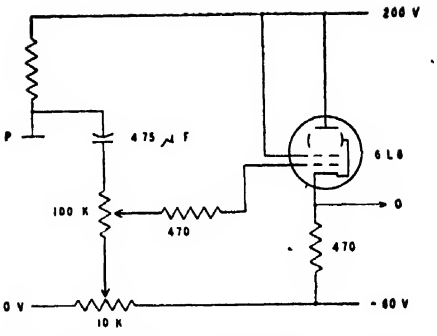


FIG. 2. Output stage of square-wave generator. *P*. Plate of clipper stage. *O*. Output to pentode current-control tubes.

longitudinal voltage gradient. These variations make it inconvenient to use a constant-voltage probe for recording arc characteristics throughout a cycle of modulation. It is much simpler to obtain these records with the aid of a probe operated at constant current, or a pair of these probes. Probe current can be held constant by use of the floating-pentode circuit shown in Fig. 3.

If the electrons in a discharge have a Maxwellian distribution of velocities, the usual plot of the logarithm of the electron current to a probe against the probe voltage has a straight-line portion[1] whose slope is proportional to the reciprocal of the electron temperature T_e. If the probe current i_P is at least 10 percent of the current i_s drawn when the probe is at space potential,† then the electron current to the probe is many times larger than the positive-ion current. In this case, the probe voltage V_P with respect to space potential is approximately equal to $(T_e/5040)\log(i_P/i_s)$. If two probes of the same area are located in regions of the discharge having the same average potential, but are operated at different currents which are on the straight-line portion of the characteristic, then the voltage between the probes is proportional to T_e. Thus two constant-current probes can be used to get a continuous record of T_e during the cycle of modulation described above, or during any transient for which the variations in i_s and T_e are small enough to permit both values of probe current to remain on the straight-line portion of the characteristic. When i_s does not vary by more than a factor of ten, the constant-current technique is more convenient than the floating-double-probe technique described by Johnson and Malter,[2] since the former gives T_e directly on a single oscillogram. (Ordinarily, two oscillograms are taken, with the probes interchanged to eliminate errors due to contact differences of potential.) One could cover a larger range of i_s by using two or more sets of probe currents, provided that one used only the portion of each record corresponding to the straight-line part of the probe characteristic. However, even this part of each oscillogram might not be a true record of T_e if the oscillograph amplifier were overloaded at any

* An article in which the dynamic characteristics of a discharge are demonstrated will be submitted by one of us (B.T.B.) for publication in the Physical Review.

† The average potential of adjacent regions of the discharge.

[2] E. O. Johnson and L. Malter, Phys. Rev. 76, 1411 (1949).

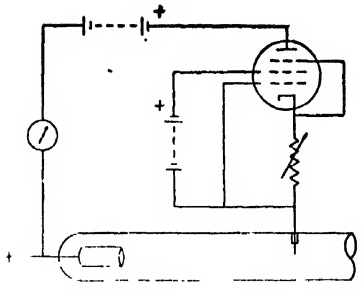


FIG. 3. Circuit used for holding probe current constant.

time during the cycle. Recovery from an overload may not be complete until several milliseconds after the input signal returns to the normal range.

If the distribution of electron velocities is non-Maxwellian, the probe characteristics may have no straight-line portion. In this case, the voltage between two constant-current probes measures the slope of a chord drawn between two points on the characteristic. The apparent T_e indicated by this slope would depend on the values chosen for the two probe currents. To find how the velocity distribution varied during a transient one would have to take a series of oscillograms—using several pairs of values for the probe currents, distributed over the probe characteristic.

Constant-current probes also are useful for measuring the difference of potential between two points in a discharge. If T_e is the same at the two points, one operates the probes at currents which are the same fraction of the respective values of i_s. This procedure allows one to obtain a continuous record of longitudinal voltage gradient, or of the radial fall in potential between the axis of the tube and the inner edge of the wall sheath.

A constant-current probe is useful as an auxiliary in two other types of probe measurement. If the current is made a little less than the minimum value of i_s, the probe will follow roughly the variations in space potential in the adjacent region of the discharge. (When i_s is at its maximum value, the probe may be 2–3 volts negative with respect to space potential.) A nearby probe made sufficiently negative with respect to the auxiliary one will collect only positive ions. By inserting a suitable resistance between the bias battery and the second probe, one can obtain a continuous record of positive ion current during a cycle of modulation.

Figure 4 shows an elaboration of this procedure which is useful for obtaining records of probe current near space potential. The cathode follower shown in this circuit allows one to bias the measuring probe with respect to the auxiliary one, and to draw whatever probe current is appropriate with this bias. In case a single 6L6 output tube does not furnish sufficient current, two or three of them are used in parallel. Two stages of direct-coupled amplification make it possible to cover a wide range of probe current with only a small change in the voltage between the input grid and the common cathode lead of the amplifier-cathode-follower circuit. Thus the voltage of the second probe follows quite closely the voltage variations of the auxiliary probe.

Using this cathode-follower, oscillograms of probe current can be taken with various bias voltages. If the latter are so chosen that probe potentials above and below space potential are obtained, then one can find i_s by making the usual semilogarithmic plot.[1] From the measured values of i_s and T_e, one can compute the electron concentration in the adjacent region of the discharge. (See reference 2, Eq. (44).)

AMPLIFIERS FOR PROBE MEASUREMENTS

In the probe measurements described above, both leads to the oscillograph amplifier have varying potentials during a cycle of modulation. One must use either a differential amplifier or an amplifier whose power supply is ungrounded. In the latter case, one must make sure that the capacitance associated with the negative lead of the supply does not cause serious errors. With a rectifier type of supply, the principal source of unwanted capacitance may be that between the primary and the secondary windings of the power transformer. In a small amplifier, this capacitance is likely to be of the order of 1×10^{-9} farads. Its effect would be most marked if the space potential at the probe suddenly shifted in the positive direction, because this would reduce the probe current. The current charging the stray capacitance in the amplifier-probe circuit could not exceed the current for which the floating pentode (Fig. 3) connected to the probe was set. If this were 10 milliamperes and if space potential shifted 10 volts, a capacitance of 1×10^{-9} farads would require only one microsecond to charge completely, if the full current were available for this time. Actually, the required time would be longer because the probe would start drawing current as its voltage approached its equilibrium value; in a typical case the adjustment would be 98 percent complete in 1.25 microseconds. Thus an amplifier with single-ended input may be entirely satisfactory for instantaneous probe measurements such as we have described, provided that the negative return lead of the

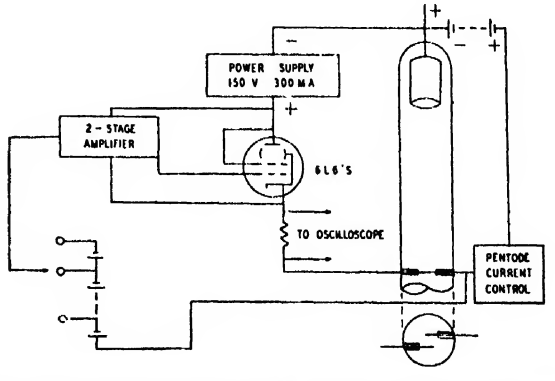


FIG. 4. Circuit for recording probe current at voltages near space potential, in order to determine electron concentration.

power supply is connected to a probe from which a current of several milliamperes is drawn.

It is advantageous to use a direct-coupled amplifier for probe measurements, since the d.c. components of voltage are needed. The frequency response should be essentially constant from 0 to at least 100,000 cycles per second. An input resistance of about 0.2 megohms and a maximum sensitivity such that two volts gives full-scale deflection have been satisfactory for our conditions of measurement (arc current >0.2 amp.). If the arc current were less than 0.1 amp., a more sensitive amplifier might be needed in order to get a reliable record of positive ion current.

RECORDING SPECTRAL INTENSITIES

To record the intensity of a spectral line or band during a cycle of modulation, one may need to use a photo-multiplier tube. An output current of the order of 10 μamp from the multiplier tube would deflect the oscilloscope spot the full height of the screen, if an amplifier similar to the one described in the preceding paragraph were used.

If the line or band to be measured is fairly strong and can be isolated with a filter, an ordinary photo-tube may suffice. In this case, it probably will be necessary to add a direct-coupled preamplifier to the amplifier used for probe measurements. To maintain adequate speed of response, the input resistance of the preamplifier ordinarily should be made no larger than 0.2 megohms. To insure sufficient stability of the zero-input trace, a d.c. heater supply should be provided for the preamplifier.

DISCUSSION

The circuits described in this article were developed for use with a mercury-vapor discharge operated in the presence of rare gas at a pressure of a few millimeters. In this case, the transient effects accompanying a sudden change in current may last several milliseconds. If the rare gas had not been present, the transients obtained with a mercury-vapor pressure of a few microns would have lasted only a few microseconds.[4] Then a 100-fold increase in both the frequency of modulation and in the speed of response of the measuring circuits would have been needed in order to get satisfactory oscillograms. This would not be possible unless the stray capacitance were kept to a minimum by use of a suitable differential amplifier and by other precautions.

Even with a gas pressure of a few millimeters, stray capacitance will be troublesome if the arc current is quite low (50 milliamperes, or less). In this case, probe currents and photo-currents will be so small that the resistors used for measuring them will have relatively high values. Then the spurious effects introduced by stray capacitance will have time constants of the same order of magnitude as those of the transients one wishes to record.

When a differential amplifier is used for probe measurements during a cycle of modulation, care must be taken to keep the amplifier operating within its linear range. Because of variations in space potential, the voltage difference between the constant-current probe and the anode or the cathode may vary over a wide range—perhaps several times as much as the voltage one wishes to measure. To offset this, one must either (1) connect the amplifier ground to a probe near to the one, or pair, with which measurements are being made, or (2) make the linear range of plate voltage for the output tubes several times the maximum voltage to be applied between the deflection plates of the oscilloscope.

SUMMARY

In studying the dynamic characteristics of a low-pressure discharge, we have found the following procedures and equipment particularly useful.

1. Pentodes in series with the discharge tube provide a convenient means of making square-wave changes in arc current.

2. A floating, self-biased pentode in series with a probe permits one to hold probe-current constant in spite of changes in the voltage gradient in the discharge. Pairs of constant-current probes can be used to get instantaneous records of electron temperature or of the voltage difference between two points in the discharge.

3. A constant-current probe is a useful auxiliary in making records of positive-ion current during a transient.

4. For measuring variations in ion concentration when the arc current is modulated, one can use a probe biased positive by various amounts with respect to a constant-current probe. The bias voltage is applied by use of a cathode-follower with two stages of direct-coupled amplification.

5. The oscilloscope amplifier used for recording probe characteristics must have both input terminals floating. An amplifier with single-ended input and ungrounded power supply may be satisfactory, but a differential amplifier is required if the arc current is relatively low or if the transients to be recorded are of very short duration.

6. Amplifiers for recording discharge characteristics should be direct-coupled.

[4] Z. P. Galkena and V. L. Granovsky, J. Tech. Phys. U.S.S.R. 18, 585 (1948).

On the Radiation Patterns of Dielectric Rods of Circular Cross Section—the $TM_{0,1}$ Mode*†

C. W. Horton, F. C. Karal, Jr., and C. M. McKinney
Defense Research Laboratory, The University of Texas, Austin, Texas
(Received July 12, 1950)

The radiation patterns have been measured for five dielectric rods of circular cross section which were excited in the $TM_{0,1}$ mode. These rods were each $0.870\lambda_0$ in diameter, and the lengths were $2\lambda_0$, $4\lambda_0$, $6\lambda_0$, $8\lambda_0$, and $10\lambda_0$. The theoretical radiation patterns have been computed by means of equivalent surface electric and magnetic currents. Excellent agreement is obtained when it is assumed that the diameter of the surface on which these currents are distributed is 0.65 of the diameter of the dielectric rod. Representative experimental and theoretical patterns are shown for the full 360°.

INTRODUCTION

ONE of the authors (C.W.H.) has developed a theory[1,2] of the radiation patterns of dielectric rods based on the radiation theorems of Schelkunoff. This theory was expanded and supplemented by experimental measurements in two later papers[3,4] of which this is a continuation. This early work was applied to dielectric rods of rectangular cross section and, consequently, suffered from a lack of knowledge of the electromagnetic field components in and on the surface of the rod. In order to minimize this difficulty, the theory has been applied to rods of circular cross section which are excited in the transverse magnetic mode with radial symmetry.

The theory developed in Part I can be applied directly to this problem. It is assumed that the contributions to the surface integral over Σ arise only from the cylindrical surface $S(r=a, 0\leqslant z\leqslant l)$ shown in Fig. 1 and, to a negligible amount, from the end cap $C(0\leqslant r\leqslant a, z=l)$. One would expect that the radius a of the surface S should be equal to the radius b of the dielectric rod. it is found, on the contrary, that the agreement with experimental data is significantly improved if the radius of S is smaller than that of the rod.

In lieu of a rigorous solution to the radiation problem, it is assumed that the fields interior to and on the surface of the dielectric rod are the same as those on an infinitely long dielectric rod of the same material and diameter. In order to test the validity of this assumption, the electric field components over the surface of the radiating rod were measured with a probe.

Since the field in a radiating rod of finite length cannot be the same as that in an infinitely long rod, two factors are introduced whose variation can be used to achieve agreement between theory and experiment. One of these factors is the radius of the surface S over which the equivalent electric and magnetic currents are distributed. The other factor arises from the assumption that the amplitude of the field components is a function of the distance along the rod.

THEORETICAL DEVELOPMENT

The nature of the waves propagated along a dielectric rod of circular cross section is discussed by Stratton.[5] He shows that a dielectric rod will propagate a transverse magnetic wave of radial symmetry, provided the frequency is greater than some minimum value. It can be shown from Stratton's discussion that in the case of a nonmagnetic and lossless dielectric rod surrounded by free space, the field components inside and on the surface of the dielectric rod are given by

$$\left.\begin{aligned} E_r &= -iE_0[n_a/(n^2-n_a^2)^{\frac{1}{2}}]J_1(\lambda_1 r)\exp(ikn_a z-i\omega t)\\ E_t &= 0\\ E_z &= E_0 J_0(\lambda r)\exp(ikn_a z-i\omega t)\end{aligned}\right\}, \quad (1)$$

$$\left.\begin{aligned} H_r &= 0\\ H_t &= -i(E_0/Z_0)[n^2/(n^2-n_a^2)^{\frac{1}{2}}]J_1(\lambda_1 r)\\ &\qquad \times\exp(ikn_a z-i\omega t)\\ H_z &= 0\end{aligned}\right\}. \quad (2)$$

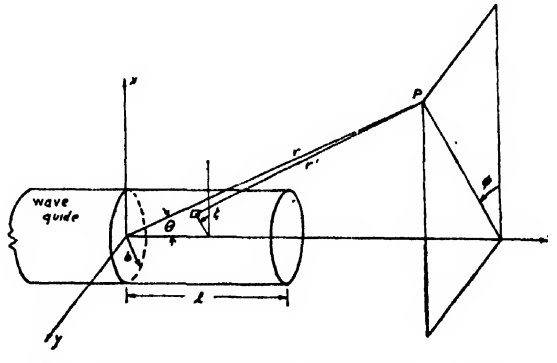


FIG. 1. The orientation of the coordinate systems with respect to the dielectric rod.

* The work described in this paper was done at the Defense Research Laboratory under the sponsorship of the Bureau of Ordnance, Navy Department.

† This paper was read at the 1950 Annual Meeting of the American Physical Society, Columbia University, New York.

[1] C. W. Horton, "On the dielectric rod as an antenna," Report No. CM-272, Defense Research Laboratory, The University of Texas, July 9, 1946.

[2] C. W. Horton, "On the dielectric rod as an antenna, II," Report No. CF-481, Defense Research Laboratory, The University of Texas, November 8, 1946.

[3] R. B. Watson and C. W. Horton, J. App. Phys. 19, 661 (1948). This will be referred to as Paper A.

[4] R. B. Watson and C. W. Horton, J. App. Phys. 19, 836 (1948). This will be referred to as Paper B.

[5] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1945), pp. 524–27.

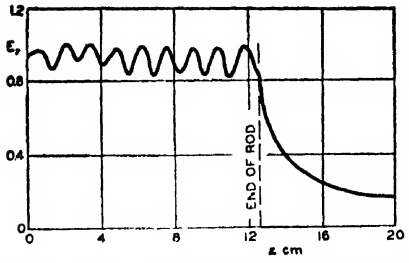


FIG. 2. The variation of the electric field component E_r with the distance along a radiating dielectric rod $4\lambda_0$ long. The mouth of the wave guide is at $l=0$.

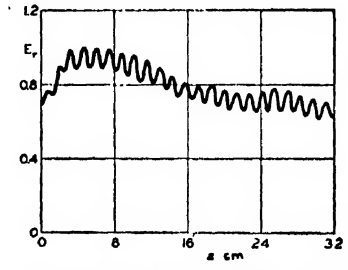


FIG. 3. The variation of the electric field component E_r with the distance along a radiating dielectric rod $10\lambda_0$ long. The mouth of the wave guide is at $l=0$.

Here, n=index of refraction of the rod material, n_a=apparent index of refraction, $k=2\pi/\lambda_0=2\pi/$free space wave-length, Z_0=impedance of free space, E_0=amplitude factor, ω=angular frequency, b=radius of the rod, and $\lambda_1=k(n^2-n_a^2)^{\frac{1}{2}}$. The apparent index of refraction, which is the ratio of the free space wave-length and the guided wave-length, is a function of the index of refraction n, the radius of the rod b, and the order of the radial mode. The value of n_a for any rod may be measured experimentally, or it may be computed from the characteristic equation.[6]

The radiation through the end C has been computed by the same method used earlier in a study of horn patterns.[7] If the dielectric rod is longer than one guided wave-length, the error introduced by neglecting the end C is not greater than 1 db at $\theta=10°$ and 1.5 db at $\theta=90°$. As the length of the rod is increased, the error caused by neglecting the end fluctuates but does not decrease because of the periodicity of the field along the rod.

When the radiation through the end C of the dielectric rod is neglected, the radiation pattern can be computed by a straightforward evaluation of a surface integral over S. The analysis is so similar to that given in Papers A and B that only the final radiation formulas are given. The integral over S reduces to the product of an integral I_1 around the circumference of S and an integral I_2 along the length of S.

In Papers A and B the possibility of radiation from standing waves as well as traveling waves was considered. The measurements of the surface fields shown in Figs. 2 and 3 show that most of the energy is found in the traveling wave. Thus the analysis of Paper B is relevant and one has for I_2

$$I_2=\frac{1}{2}\int_0^l \omega(z)\exp(iqkz)dz, \qquad (3)$$

where $\omega(z)$=a shading function arbitrarily introduced to allow for the termination of the rod, $q=n_a-\cos\theta$, and θ=angle between the axis of the rod and a line to the field point (see Fig. 1).

For the transverse magnetic wave whose components are given by (1) and (2), the components of the radiation field at a point P in the far zone are given by

$$E_r=0,$$

$$E_\theta=-(ika/2\pi r)(\exp ikr)I_2 \times\int_0^{2\pi}[Z_0\bar{H}_\zeta\sin\theta-\bar{E}_z\cos(\phi-\zeta)] \times\exp[-ika\sin\theta\cos(\phi-\zeta)]d\zeta, \qquad (4)$$

$$E_\phi=0.$$

Here the barred quantities $\bar{H}_\zeta$ and $\bar{E}_z$ are the same as the corresponding unbarred quantities except that the factor $\exp(ikn_az)$ has been split off. The components of H are easily found, since this is a transverse wave.

Since it has been found necessary to introduce a distinction between the physical radius of the rod and the radius of the equivalent surface S, one would expect

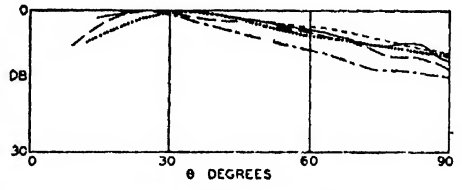


FIG. 4. The envelopes of the experimental patterns *versus* angle. — — — — $l=10\lambda_0$. ——— $l=8\lambda_0$. — — — — — — $l=6\lambda_0$. — — — — — — — — $l=4\lambda_0$. . . . . $l=2\lambda_0$.

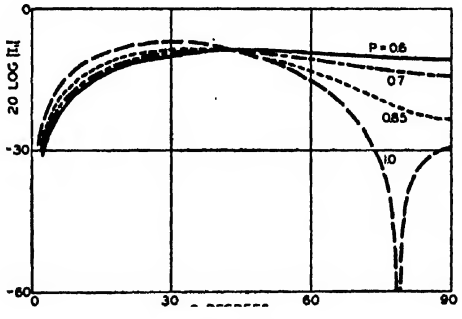


FIG. 5. The angular variation of $20\log|I_1|$ for various values of the ratio P of the radius of the surface S to the radius of the dielectric rod. The field components E_z and H_ζ are evaluated on the surface S.

[6] See reference 5, Eq. (9), p. 526.
[7] C. W. Horton, Proc. I.R.E. 37, 744 (1949).

that field components H_ξ and E_z should be evaluated on the surface S. This has been done and good agreement was obtained. However, it was necessary in this case to make the radius of the surface S smaller as the length of the rod was increased. On the other hand, if the field components were evaluated on the surface of the physical rod, a constant diameter for S is obtained for rod lengths from $2\lambda_0$ to $10\lambda_0$. Furthermore, the agreement between experimental and theoretical patterns is even better than in the first case.

When one examines (4), it is clear that the choice of the surface S on which H_ξ and E_z are evaluated is significant only in that it determines the ratio E_z/H_ξ, or, in other words, it changes the relative contribution of the magnetic currents and the electric currents to the radiation field.

When the integration is performed, one finds

$$E_\theta \propto \psi I_1 I_2, \tag{5}$$

where the constant of proportionality is immaterial for relative patterns. In Eq. (5)

$$\psi=(1/r)\exp(ikr-i\omega t),$$
$$I_1=J_0(ka\sin\theta)\sin\theta-CJ_1(ka\sin\theta), \tag{6}$$

$$C=\frac{E_z}{H_\xi}\bigg|_r=\frac{(n^2-n_a^2)^{\frac{1}{2}}}{n^2}\frac{J_0(\lambda_1 r)}{J_1(\lambda_1 r)}, \tag{7}$$

r = surface at which E_z and H_ξ are evaluated.

The integral I_2, which is defined in (3), will be discussed later. The envelope of the radiation pattern is determined primarily by I_1, while the lobe structure arises from I_2.

EXPERIMENTAL OBSERVATIONS

In order to test the equations derived in the preceding section, a series of rods of constant radius were constructed of Lucite. The diameter of each of these rods was $0.870\lambda_0$, but the lengths of the rods were $2\lambda_0$, $4\lambda_0$, $6\lambda_0$, $8\lambda_0$, and $10\lambda_0$. The rods fitted snugly in the wave guide, and the ends of the rods were terminated by a tapered section to avoid a sudden discontinuity. The index of refraction of the Lucite was measured[8] by a transmission method and a reflection method and found to be 1.60 ± 0.01.

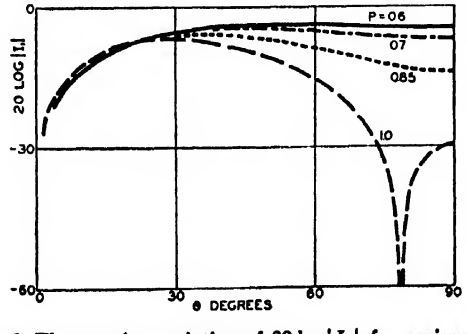


FIG. 6. The angular variation of $20\log|I_1|$ for various values of the ratio P of the radius of the surface S to the radius of the dielectric rod. The field components E_z and H_ξ are evaluated on the surface of the dielectric rod.

TABLE I. The value of the parameters d and D that yield best agreement with the experimental patterns. $b=0.435\lambda_0$; $P=0.65$.

Length l/λ_0	d	D
2	2.0	−0.5
4	1.11	−0.1
6	1.0	0
8	0.5	1.0
10	0.3	2.33

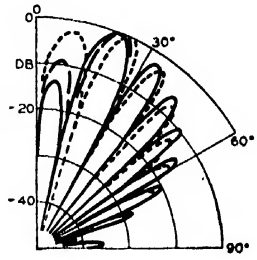


FIG. 7. An illustration of the effect of the shading factor d on the theoretical pattern. —— $d=1.0$; — — — $d=0.5$; - - - - $d=0.2$. Length of rod $=6\lambda_0$. $P=1.0$.

In the experimental pattern measurements, the dielectric rods were used as transmitters. The power source consisted of a type 723 a/b velocity modulated tube operated at a frequency of 9375 m c.p.s. or a wavelength of 3.20 cm. The $TM_{0,1}$ mode was obtained with a mode transducer. A motor-driven turntable was used to rotate the transmitting antenna through 360° in azimuth. The receiving antenna was 40 feet from the transmitting antenna, and both antennas were mounted approximately 12 feet above the roof of a three-story building. The output of the receiver was recorded on an Esterline-Angus recording milliammeter.

Figures 2 and 3 show the dependence of E_r at the surface of the rod on the distance along the rod for the rods of length $4\lambda_0$ and $10\lambda_0$, respectively. These measurements were made with a small probe while the antenna was radiating. These measurements and similar ones on the other rods of the series show low standing wave ratios. They also show that the apparent index of refraction for the series of rods is 1.163 ± 0.008. The value obtained from long rods used as wave guides is 1.175 ± 0.008. It is interesting to note that the transition from the field in the metallic guide to the field in the dielectric guide appears to be complete in a distance of approximately one or two guide wave-lengths.

THE CONTRIBUTION OF THE INTEGRAL I_1

The envelopes of the experimental patterns throughout the first 90° of the five rods are shown in Fig. 4. It appears that the envelope oscillates about a mean value that is independent of the length. Figure 5 shows a series

[8] C. G. Montgomery, *Technique of Microwave Measurements* (McGraw-Hill Book Company, Inc., New York, 1947), Chapter 10.

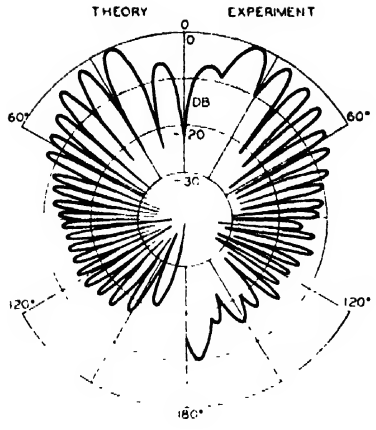


FIG. 8. The experimental and theoretical patterns of a Lucite rod $10\lambda_0$ long and $0.87\lambda_0$ in diameter.

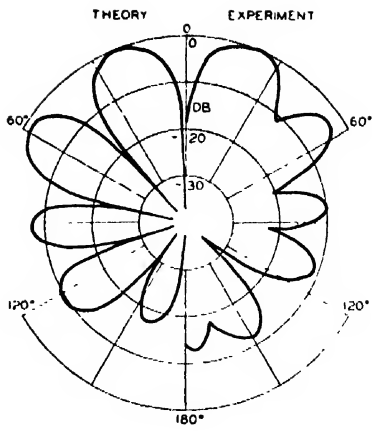


FIG. 10. The experimental and theoretical patterns of a Lucite rod $2\lambda_0$ long and $0.87\lambda_0$ in diameter.

of curves of $20\log|I_1|$ for various values of P, the ratio of the radius of the surface S to the radius of the dielectric rod; i.e., $P=a/b$. In calculating these curves, the field components E_z and H_t are evaluated on the surface S. The curves are plotted to 90° only, because the theoretical curve for I_1 is symmetric about 90°. Also, beyond 90° the influence of I_2 is so great that the behavior of I_1 alone is not significant.

Figure 5 shows that the pattern depends on the radius of the surface S qualitatively in the same way that any radiation pattern depends on a physical dimension of the source. Further, it is evident that agreement with the experimental curves shown in Fig. 4 will not be obtained unless P is less than one. In particular, it was found that if the field components were evaluated on S, the optimum values of P depended on the length as follows: $P=0.85$, 0.70, 0.65, 0.625, 0.60 for $l/\lambda_0=2$, 4, 6, 8, 10, respectively.

Figure 6 shows a plot of $20\log|I_1|$ for various values of P when the field components E_z and H_t are evaluated on the surface of the physical rod. A comparison of Figs. 5 and 6 shows that the variation of $|I_1|$ with the angle θ is less when the ratio E_z/H_t is evaluated on the surface of the physical rod than when it is evaluated on the surface S.

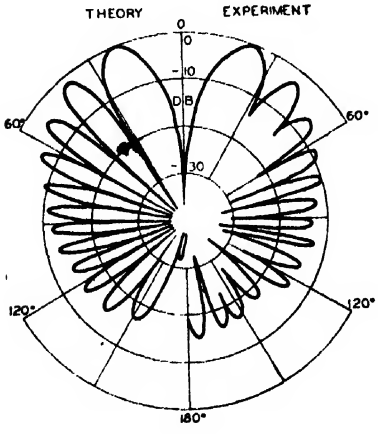


FIG. 9. The experimental and theoretical patterns of a Lucite rod $6\lambda_0$ long and $0.87\lambda_0$ in diameter.

THE VARIATION OF AMPLITUDE ALONG THE LENGTH OF THE ROD

Before one can compute theoretical patterns for comparison with theory, it is necessary to evaluate the integral I_2 defined in (3). Undoubtedly, the best method of procedure would be to determine the shading function $\omega(z)$ for each individual rod from surface measurements such as those shown in Figs. 2 and 3. However, since a very large number of patterns have been measured and computed, it is more convenient to choose a standard form for the function $\omega(z)$ and to obtain agreement with the theoretical patterns by a variation of the parameters in this function.

Mueller and Tyrell[9] have found that a function $\omega(z)$ of the form

$$\omega(z)=d+(1-d)\sin(\pi z/l), \tag{8}$$

is useful in discussions of dielectric rods. This function includes as a special case, $d=1$, the case treated in Paper B. When (8) is substituted into (3), one obtains

$$I_2=\frac{ld}{2}\left|\frac{(\pi/2)D\cos\phi}{(\pi/2)^2-\phi^2}+\frac{\sin\phi}{\phi}\right|\exp(i\phi), \tag{9}$$

where

$$\phi=(kl/2)(n_a-\cos\theta),$$
$$D=(1-d)/d.$$

The constant factor $(ld/2)$ may be ignored in a study of relative patterns.

In order to illustrate the effect of d (or D) on the pattern, three theoretical patterns are shown in Fig. 7 in which only d is changed. These patterns are computed for the rod F-4 when the radius of the surface S is equal to the radius of the dielectric rod. It is seen that the first lobe is very sensitive to the choice of d, while the third and higher lobes are not very sensitive. This conclusion is borne out by a study of the extrema of (9).

The first major lobes of the patterns result from the interaction of I_1 and I_2 and not from a maximum in either one separately. The remaining lobes are caused

[9] G. E. Mueller and W. A. Tyrell, *Bell Syst. Tech. J.* **26**, 837 (1947).

by extrema in I_2, although their exact positions are influenced by the behavior of I_1. This interaction is significant in the rods $2\lambda_0$ and $4\lambda_0$ long, but when the rod is $6\lambda_0$ or more in length, the angular positions of the lobes agree with the extrema of $(\sin\phi)/\phi$ within the limits of experimental error.

THE THEORETICAL PATTERNS

When the values of d (or D) are adjusted to give best agreement with the experimental data, one finds the values listed in Table I. The values are based on the assumption that the electromagnetic field components are evaluated on the surface of the rod but that the radius of the surface S is only 0.65 of the radius of the rod.

Theoretical patterns for all the rods were computed by means of (5) on the basis of the assumptions just mentioned and for the values of d shown in Table I. The excellence of the agreement is illustrated in Fig. 8 which shows on the right the experimental pattern of a rod $10\lambda_0$ long and on the left the theoretical pattern computed by means of (5). Figures 9 and 10 show similar patterns for rods $6\lambda_0$ and $2\lambda_0$ long, respectively.

SUMMARY

In view of the accurate knowledge that one has of the distribution of the field components in a dielectric rod of circular cross section, the formulas developed in this paper apply to any length of rod unless the rod is so short that the effect of the end cannot be neglected. This is in pleasing contrast to the work in Paper A in which the theory failed for rods longer than $6\lambda_0$. The difficulty with the height of the minor lobe encountered in Fig. 6 of Paper A has vanished also.

Precision Determination of the Lattice Constants of Zinc Oxide

R. B. Heller, J. McGannon,† and A. H. Weber
Physics Department, Saint Louis University, St. Louis, Missouri
(Received June 12, 1950)

A total of 350 back-reflection ZnO powder x-ray diffraction lines on 25 photographic films have been measured and analysed by the Cohen analytical least squares extrapolation method to yield $a=(3.2495\pm0.0000_3)$A, $c=(5.2069\pm0.0001_1)$A; where the stated errors are the probable errors. The statistical reliability of the results is emphasized.

In the measurement of the Debye-Scherrer ring radii the accuracy obtained with a metric scale and indicator (employing no optical lens system) and with a dividing engine equipped with a microscope are compared and the former simpler method is found more reliable. It is concluded that the use of precision measuring instruments for observing diffraction ring radii are unnecessary and even undesirable apparently because the magnifying power of the optical system of precision micrometers leads to uncertainties in settings on photographic lines.

THE Cohen analytical extrapolation method[1] has been applied to twenty-five powder photographs of ZnO in the back-reflection region using Cu $K\alpha$ x-radiation employing the Straumanis technique.

The specimen was granulated ZnO powder* of purity better than 99.85 percent. The powder was ground and sifted through a 200-mesh screen to insure some uniformity of fragment size. It was packed into an enamel tube (0.025-in. diameter, 0.5-in. length). This cylindrical specimen holder is the coating of magnet wire,** sleeved off by stretching the wire. The specimen was then carefully centered in a Debye camera (57.296-mm radius) and exposed to x-rays. The Debye-Scherrer rings were measured to 0.005 in. with a metric scale.

From each of the twenty-five photographs data were recorded for a total of fourteen lines, using reflections of the $K\alpha_2$ and $K\alpha_1$ radiations. The procedure was to convert all $K\alpha_2$ reflections to the equivalent $K\alpha_1$ reflection by multiplying $\sin^2\theta$ by $(\lambda K\alpha_1/\lambda K\alpha_2)^2$. The values[2] of Cu wave-lengths ($\lambda K\alpha_1=1.54050$A, $\lambda K\alpha_2=1.54434$A) recommended by the American Society for X-ray and Electron Diffraction were used.

Data taken from one of the films are shown in Table I. $\Phi'=(\pi-2\theta/2)$ is one-half the back-reflection angle and θ is the Bragg angle. The results for the twenty-five films are shown in Table II. The stated error is the probable error of all the determinations.

To determine the effect of measuring a greater number of significant figures in the Debye-Scherrer radii, a dividing engine reading to 0.00025 cm was employed for remeasuring of the same twenty-five films. Because of the limited length of the dividing engine scale, the Straumanis method could not be employed. Instead, a constant R, obtained by calibrating the camera with NaCl, was used. The application of the Cohen method yields: $a=(3.24892\pm0.00008)$A and $c=(5.20489\pm0.00037)$A. The spread for a is 0.00244A and that for c is 0.01245A. Note that the spread here is greater than that of the results shown in Table II.

† Now at Creighton University, Omaha, Nebraska.
[1] M. U. Cohen, Rev. Sci. Inst. 6, 68 (1935).
* Obtained from Mallinkrodt Chemical Works, St. Louis, Mo.
** Obtained from Westinghouse Electric Corporation, East Pittsburgh, Pa.
[2] E. A. Wood, Phys. Rev. 72, 436 (1947).

TABLE I. Detailed data for film 5 using the first method (Straumanis method, employing simple measuring scale).

s^* (cm)	Φ' (radian)	θ (radian)	hkl	α	γ	$(\sin^2\theta_{\text{assumed}}-\sin^2\theta_{\text{exp}})^{**}=\gamma\times10^{-8}$	δ
1.061	0.0926	1.4782	222	12	4	0.99117948−0.98652603=0.465345	0.033738
1.330	0.1161	1.4547	222	12	4	0.99117948−0.98658209=0.459739	0.052951
1.672	0.1460	1.4248	310	13	0	0.97840682−0.97397438=0.443244	0.082452
1.868	0.1631	1.4077	310	13	0	0.97840682−0.97363461=0.477221	0.102681
3.619	0.3159	1.2549	220	12	0	0.90314476−0.89899759=0.414717	0.347067
3.710	0.3239	1.2469	220	12	0	0.90314476−0.89870885=0.443591	0.364125
4.091	0.3572	1.2136	214	7	16	0.87897334−0.87338604=0.558730	0.427101
4.148	0.3621	1.2087	214	7	16	0.87897334−0.87451729=0.445605	0.438947
4.279	0.3734	1.1974	106	1	36	0.86757458−0.86262977=0.494481	0.459141
4.356	0.3803	1.1905	106	1	36	0.86757458−0.86221381=0.536077	0.475205
4.534	0.3959	1.1749	205	4	25	0.85126527−0.84705720=0.420807	0.503879
4.607	0.4022	1.1686	205	4	25	0.85126527−0.84590896=0.535631	0.518990
5.797	0.5062	1.0646	302	9	4	0.76539329−0.76111732=0.427597	0.715704
5.844	0.5103	1.0605	302	9	4	0.76539329−0.76143044=0.396285	0.726616

$\alpha=(h^2-k^2-hk)$, $\gamma=l^2$, $\delta=\sin^2 2\theta_{\text{exp}}$.

Employing the Cohen‡ analysis:

$$\begin{aligned}1208\,\Delta A_1+\ 664\,\Delta C+36.05034\,D&=0.5266834,\\ 664\,\Delta A_1+\ 4418\,\Delta C+79.1810\ D&=0.840763,\\ 36.05034\,\Delta A_1+79.1810\,\Delta C+2.649494\,D&=0.02431065.\end{aligned}$$

$$\Delta A_1=0.000345186,\quad A_1'=0.075262063,\quad A_1=A_1'-\Delta A_1';$$
$$\Delta C=0.00012523,\quad C'=0.02200868,\quad C=C'-\Delta C.$$

Finally,

$$a=\frac{\lambda}{(3A_1)^{\frac{1}{2}}}=3.2495\text{A},\quad c=\frac{\lambda}{(4C)^{\frac{1}{2}}}=5.2068\text{A}.$$

* Each s value is the mean of five readings.
** Numerical values from: Federal Works Agency, Work Projects Administration for the City of New York, *Tables of Sines and Cosines*, 1940, which permit accurate interpolation for eight figures in sine value for five figures in θ.
‡ See reference 1.

TABLE II. Lattice constants of ZnO. Results from twenty-five films using the first method (Straumanis method, employing simple measuring scale). $\lambda K\alpha_1=1.54050$A; $\lambda K\alpha_2=1.54434$A.

a	c
3.2495A	5.2070A
3.2497	5.2072
3.2496	5.2076
3.2497	5.2088
3.2495	5.2068
3.2495	5.2063
3.2494	5.2077
3.2494	5.2066
3.2494	5.2060
3.2494	5.2075
3.2492	5.2064
3.2493	5.2065
3.2493	5.2058
3.2493	5.2052
3.2495	5.2071
3.2495	5.2060
3.2494	5.2063
3.2493	5.2069
3.2496	5.2066
3.2496	5.2082
3.2495	5.2067
3.2498	5.2083
3.2496	5.2074
3.2496	5.2065
3.2494	5.2063
(mean) 3.2495A	(mean) 5.2069A

$a=(3.2495\pm0.00002)$A,
$c=(5.2069\pm0.00011)$A.

To determine whether this larger spread is due to employing a constant R (not using the individual calibration of the Straumanis method) or to greater uncertainties in the dividing engine readings, five of the twenty-five films were analyzed in the following manner. The Debye-Scherrer radii measurements of the second method (employing dividing engine) were combined with the Straumanis calibration for camera radius data available from the first method. The results of this procedure are $a=3.24877$A with a spread of 0.00127A and $c=5.20549$A with a spread of 0.00414A. The values of a and c determined from the same five films using the second method (dividing engine and constant R) show a spread in a of 0.00125A and a spread in c of 0.00415A. These almost identical results show definitely that the uncertainties in setting the dividing engine, not the small variation in R, are responsible for the original larger spread of the second method. Therefore, a practical conclusion is that the use of a precision measuring instrument, such as the dividing engine, for measuring Debye-Scherrer ring radii is not necessary and is even undesirable apparently because the magnifying power of the optical system of precision micrometers leads to uncertainties in settings on the photographic lines.

Bunn's values[3] for a and c (obtained by the Bradley-Jay graphical extrapolation method, not the Cohen analytical method) corrected for the wave-lengths here employed are: $a=(3.2492\pm0.0001)$A and $c=(5.2054\pm0.0003)$A. It is seen that the present Table II results are somewhat larger than those of Bunn. As originally observed and explained by Bunn,[3] in the present work the error in c is larger than the error in a. It is pointed out that in the present work a total of 350 back-reflection lines were measured and therefore the statistical reliability should be good.

[3] C. W. Bunn, Proc. Roy. Soc. 47, 835 (1935).

The Virtual Mass of a Sphere Entering Water Vertically*

ALBERT MAY AND JEAN C. WOODHULL
Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland
(Received June 23, 1950)

The manner in which a sphere behaves when it enters the water from air depends on the rate at which energy is taken from it, and most of this energy is utilized in setting water into motion. The virtual mass is a partial measure of this motion. An investigation of virtual mass for spheres shortly after vertical water entry is described. The spheres used had specific gravities between 1.06 and 16.77 and impact velocities ranged from 21 to 103 ft./sec. The value of the virtual-mass coefficient was found to be much smaller than had previously been estimated, an average value of 0.08 being obtained. The method used recognizes the dependence of C_D on sphere speed and depth, and makes comparisons only where these parameters have the same value.

INTRODUCTION

THE concept of virtual mass is well known[1] in its application to the motion of a body completely immersed in a fluid. It arises from the fact that when a body moves through a fluid a quantity of fluid shares in the motion of the body. When the body is accelerated, this virtual mass (or "apparent additional," "induced," or "dragged" mass) must be accelerated also, and the body accordingly behaves approximately as though its mass were increased by an amount equal to the virtual mass.

The virtual-mass concept is useful, but not necessary, in the mathematical treatment of such problems as the accelerated motion of a completely immersed body. When this concept is used the assumption is made that the drag force, dependent on the speed but not on the acceleration, acts on both the moving body and on the entrained mass of water. In other words, the entrained water is treated as if it were a part of the moving body and the effective forces are those which act on their combined masses. This viewpoint has been used in the present report. Alternatively, attention could be focused on the actual forces (e.g., drag) acting on the sphere alone. For accelerated motion the magnitude of such forces would depend on the amount of water entrained by the body. Hence from this second viewpoint the drag coefficient is a function of the acceleration of the body, but the drag forces act totally on the sphere itself.

If potential flow is assumed, computation for a completely immersed sphere moving through a fluid shows the increase of mass to be one-half the mass of the displaced fluid.[2] This result is reasonably verified by experiment.[3]

The drag which a body experiences when it moves in a fluid is not measured directly but is deduced from the acceleration. Generally it is expressed in terms of the drag coefficient C_D which is defined by the relation,

$$\text{drag}=\tfrac{1}{2}\rho v^2 A C_D, \tag{1}$$

where ρ is the density of the fluid, and v and A are the speed and cross-sectional area of the missile.

The equation of motion of a body moving through a fluid may be written

$$-(1/v^2)dv/dt=d(1/v)/dt=\alpha, \tag{2}$$

where the retardation coefficient α is constant or variable, depending upon the conditions under which the body is moving. Obviously from Eq. (2), if $1/v$ is plotted against t, the slope of the graph will be α. To study the drag at water entry[4] high-speed motion pictures are taken of the entry, time and position data are read from the film, and instantaneous values of speed are obtained from increments of distance and time. Next a graph is drawn of reciprocal speed against time and from its slope α is obtained.

So far the analysis has been entirely kinematic. Since α is a measure of the acceleration, its value, of course, depends on the mass of the sphere, but it is uniquely determined by position-time data and does not depend on the viewpoint that is chosen in regard to the mass of the moving body, that is, whether it is taken as the mass of the body alone or this increased by the virtual mass.

The mass of a homogeneous body may be written as $V\sigma\rho$, where V is the volume of the body, σ its specific gravity relative to water, and ρ the absolute density of water. When the body is moving in water the mass of entrained water may be written $kV\rho$, where k is a proportionality constant which will be called the virtual-mass coefficient. This coefficient has also been known as the "induced mass factor" and the "relative apparent mass." The total mass accelerated is therefore $(\sigma+k)V\rho$.

From Eq. (2) the acceleration is αv^2. Hence,

$$\text{drag}=(\sigma+k)V\rho\alpha v^2. \tag{3}$$

Suppose that two spheres differing only in specific gravity are moving vertically downward at the same speed. If it is assumed that the drag force is the same

* Sponsored by the ONR.

[1] Cf. H. Bateman, "Report of the Committee on Hydrodynamics," Part IV; National Research Council Bulletin No. 84 (National Research Council, Washington, D. C., 1932).

[2] H. Lamb, *Hydrodynamics* (Dover Publications, Inc., New York, 1945), 6th Edition, p. 124.

[3] G. Cook, Phil. Mag. 39, 350 (1920).

[4] A. May and J. C. Woodhull, J. App. Phys. 19, 1109 (1948).

TABLE I. Measured specific gravities of ½-inch spheres used.

Material	Specific gravity
Polystyrene	1.06
Magnesium	1.79
Steel	7.76
Tungsten	16 77

for the two spheres, and that it acts both on the sphere itself and on the added mass of the entrained water, then

$$\text{drag}=(\sigma_1+k)V\rho\alpha_1 v^2=(\sigma_2+k)V\rho\alpha_2 v^2, \qquad (4)$$

where the subscripts 1 and 2 refer to the lighter and heavier spheres, respectively. From Eq. (4) we may write

$$k=\frac{\sigma_2\alpha_2-\sigma_1\alpha_1}{\alpha_1-\alpha_2}. \qquad (5)$$

The experiment just outlined furnishes one of the principal means of investigating virtual mass.

DETAILS OF THE METHOD

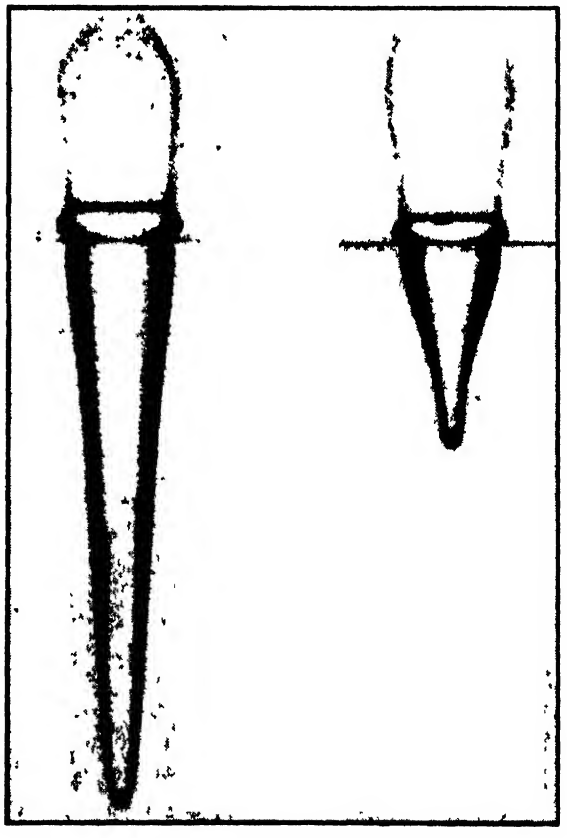
In the present paper, calculations of virtual mass will be given only for the vertical water entry of spheres and in particular for that phase which immediately follows the actual entry, during which an air-filled cavity is still attached to the sphere. Previous determinations of virtual mass for this case have been made in an attempt to systematize the empirical data of water entry but have apparently been described only in literature which was restricted in distribution for reasons of national security. A value of 0.40 or 0.45 was generally reported for k. The values of α (or C_D) from which these determinations were made were not comparable in precision with values obtainable today.

Numerous improvements in method were found desirable or necessary in the course of the present work. In the determinations of α that we reported previously,[4] values of $1/v$ were plotted against t and the slope of a least-square straight line through these points was taken as the value desired. The accuracy has now been improved sufficiently to justify replacing this straight line by a curve which yields instantaneous values of α.

When Eq. (5) is used for determining the virtual-mass coefficient it is assumed that the same drag acts on the two spheres used. Forces which are not hydrodynamic in nature should preferably be eliminated but it is sufficient that they be the same for the two spheres. In obtaining values of k from Eq. (5), we have always compared two spheres when the speed and depth of the spheres were the same and their cavities had the same pressure. Further, as will be discussed later, the contribution of weight to the total force on the sphere was removed mathematically.

Each determination of k was made from the water entry of a pair of half-inch spheres: one of steel or tungsten and the other of polystyrene or magnesium. The measured specific gravities of these spheres are given in Table I. The choice of such a pair depends on the fact that if two shots are so chosen that the entry speed of the lighter is slightly greater than that of the heavier, corresponding points can always be found in the two trajectories where the depths and speeds are the same. The choice of such conditions prevents the drags differing because of different speeds and should result in reasonably equal hydrostatic buoyant forces on the two spheres.

The effect of cavity pressure was removed by reducing the air pressure above the water, thus preventing surface closure of the cavity.[5] The cavities obtained and the considerable difference in cavity formation for light and heavy spheres can be seen in the photographs of Fig. 1. These were taken 26 milliseconds after the ½-inch tungsten and polystyrene spheres had entered the water at 63 ft./sec. The tungsten sphere had reached about three times the depth of the polystyrene sphere and was traveling about three times as fast. At shallower and equal depths the cavities are much more alike than those of Fig. 1. This is important since different



a b

FIG. 1. a. Tungsten; b. Polystyrene. Photographs of ½-inch spheres 26 milliseconds after they entered water at 63 ft./sec., showing dependence of cavity shape on sphere density.

[5] D. Gilbarg and R. A. Anderson, J. App. Phys. 19, 127 (1948).

TABLE II. Determinations of virtual-mass coefficient.

Low density sphere				High density sphere				Compared at		
Mat'l	Entry speed (ft./sec.)	α_2(in.$^{-1}$)	C_D	Mat'l	Entry speed (ft./sec.)	α_1(in.$^{-1}$)	C_D	Speed (ft./sec.)	Depth in diameters	k
Poly.	59	0.356	0.251	Steel	21	0.0527	0.273	18.1	6.4	0.11
Poly.	59	0.350	0.247	Steel	32	0.0514	0.266	29.0	3.7	0.10
Poly.	59	0.350	0.247	Steel	38	0.0487	0.252	34.2	2.6	0.03
Poly.	62	0.358	0.253	Steel	32	0.0514	0.266	29.0	3.6	0.06
Poly.	64	0.347	0.245	Steel	21	0.0527	0.273	17.9	6.6	0.14
Poly.	64	0.341	0.241	Steel	32	0.0516	0.267	28.9	3.8	0.14
Poly.	64	0.341	0.241	Steel	38	0.0506	0.262	34.2	2.8	0.11
Poly.	64	0.341	0.241	Steel	38	0.0487	0.252	34.2	2.8	0.04
Poly.	103	0.344	0.243	Steel	32	0.0520	0.269	26.6	7.1	0.13
Poly.	103	0.343	0.242	Steel	65	0.0522	0.270	59.9	2.3	0.14
Poly.	103	0.343	0.242	Steel	78	0.0524	0.271	76.6	0.9	0.15
Poly.	103	0.344	0.243	Tungsten	40	0.0231	0.258	39.1	4.7	0.07
Poly.	103	0.343	0.242	Tungsten	62	0.0245	0.274	59.5	2.3	0.15
Poly.	103	0.343	0.242	Tungsten	62	0.0248	0.277	58.7	2.4	0.16
Mg.	62	0.228	0.273	Steel	21	0.0551	0.285	15.9	12.1	0.10
Mg.	62	0.218	0.261	Steel	32	0.0522	0.270	26.3	7.7	0.09
Mg.	62	0.218	0.261	Steel	38	0.0510	0.264	31.7	6.0	0.03
Mg.	62	0.218	0.261	Steel	38	0.0495	0.256	31.7	6.0	−0.04
Mg.	95	0.247	0.296	Steel	21	0.0572	0.296	14.5	16.3	0.00
Mg.	95	0.200	0.263	Steel	32	0.0529	0.274	23.6	11.9	0.10
Mg.	95	0.221	0.264	Steel	38	0.0516	0.267	28.9	10.0	0.03
Mg.	95	0.223	0.267	Steel	38	0.0502	0.260	28.9	10.0	−0.04
Mg.	95	0.218	0.261	Steel	58	0.0522	0.270	53.4	4.5	0.08
Mg.	95	0.218	0.261	Steel	65	0.0522	0.270	58.0	3.6	0.08
Mg.	95	0.217	0.260	Steel	78	0.0524	0.271	77.0	0.8	0.11
Mg.	95	0.219	0.262	Tungsten	40	0.0232	0.259	37.9	7.8	−0.02
Mg.	95	0.218	0.261	Tungsten	62	0.0246	0.275	58.5	3.7	0.11
Mg.	95	0.218	0.261	Tungsten	62	0.0248	0.277	57.5	3.8	0.12

cavity shape means a different flow pattern and probably different drag.

In writing Eq. (2) the effect of gravity was omitted. Instead, it may be written

$$d(1/v)/dt = -g/v^2 + \alpha. \tag{6}$$

It is seen from Eq. (6) that the slope of the $1/v$ against t graph is $-g/v^2+\alpha$. The effect of gravity can be eliminated and α determined, by adding g/v^2 to the slope of the graph at any desired point. This correction of α effectively eliminates the weight which otherwise would make large and considerably different contributions to the drag of the light and heavy spheres.

RESULTS

High-speed motion pictures were taken of numerous shots of each of the spheres listed in Table I at various entry speeds from 21 to 103 ft./sec. Position-time data were read from each film, values of $1/v$ were plotted against t, and a smooth curve was drawn. By comparing each shot of a dense sphere with each suitable shot of a light one, the corresponding positions were found where the speed and depth of the spheres were the same. The slopes of the $1/v$ against t graphs were read at these points and values of α were obtained by adding g/v^2 to the slopes. Finally k was found by means of Eq. (5).

The values found for k and related information are given in Table II. These values for k range from −0.04 to 0.16. Negative values are surely meaningless but may give some indication of the accuracy attainable. It should be noted especially that the values here reported for k are considerably smaller than the values of about 0.45 previously obtained by others. A small part of this difference may be due to the low ambient pressure used in the present work. This changes the cavity shape and hence the flow pattern, and may change the virtual mass slightly.

An attempt was made to determine whether the variations of k shown in Table II were due to variations of depth or of instantaneous velocity, by plotting the k values against these quantities as coordinates in Fig. 2. Nothing significant appears in this presentation.

In studying the self-consistency of the data, the scatter of the data will appear worst if all variation of k is regarded as due to experimental error, rather than to dependence of k on experimental parameters. On this basis the data yield a mean value of 0.08 for k and a standard deviation of 0.06.

It should be noted that the average value of k obtained by use of polystyrene spheres was twice that for the magnesium spheres: 0.109 for polystyrene, 0.054 for magnesium. The reason for this is not known.

An interesting view of the virtual-mass coefficient may be obtained by a method which requires that random errors in determining the values of α be more important than actual variations in the value of α. Suppose that $1/v^2$ times the drag, in Eq. (4), is negligibly dependent on sphere density, depth and speed. Then, since the expression in Eq. (4) has the same value for either subscript 1 or 2, this subscript can be written as i,

and Eq. (4) becomes:

$$\alpha_i\sigma_i + k\alpha_i = \text{constant}. \tag{7}$$

If $\alpha_i\sigma_i$ is plotted against α_i, Eq. (7) shows that a straight line should be obtained, of slope $-k$. The data of Table II are plotted in Fig. 3, but no attempt has been made to draw in this line for the determination of k, because of the large spread in the data. Instead groups of straight lines are seen drawn transverse to the principal extension of the data. Each of these lines represents the spread of data obtained from one water entry and the individual values of α from Table II are shown as dots on these lines. To determine k it would be possible to draw straight lines which lie within the range of values plotted for all materials. They would give k values between 0.07 and 0.13.

DISCUSSION

The reasons why the value of virtual mass obtained for the cavity phase of vertical water entry is so small can best be understood by considering the significance of virtual mass as it occurs in various idealized hydrodynamic problems.

For potential flow about a completely immersed body, with no attached cavity, energy is conserved and the drag is zero. The fluid surrounding the body has at each point a speed which is proportional to that of the body, and the body is in "immediate" communication with all of the water affected by the body's motion. When the speed of the body changes, the speed of each particle of the water changes proportionately. The virtual mass is that mass of the fluid which would have a kinetic energy equal to the actual kinetic energy of the fluid if this mass had the same speed as the body, or $(\int v^2 dm)/v_s^2$, where v is the speed of the fluid at any point and v_s is the speed of the body at the same instant. For this situation all the water which is in motion contributes in some degree to the virtual mass.

Let us now determine the virtual mass by subjecting the completely immersed body to an imaginary acceleration. The speeds of all water particles will be changed proportionately and the total work done by the accelerating force will be that done in changing the speed of the combined mass; mass of the sphere plus virtual mass.

Next let the proposed experiment be repeated with two bodies that differ only in specific gravity, and arrange that the total forces are the same on the two bodies (disregarding gravity, whose influence can be removed mathematically). The accelerations will be different for the two bodies. From the accelerations the virtual mass can be calculated. The experimental tests described in this report were of this type except that cavities were attached to the bodies.

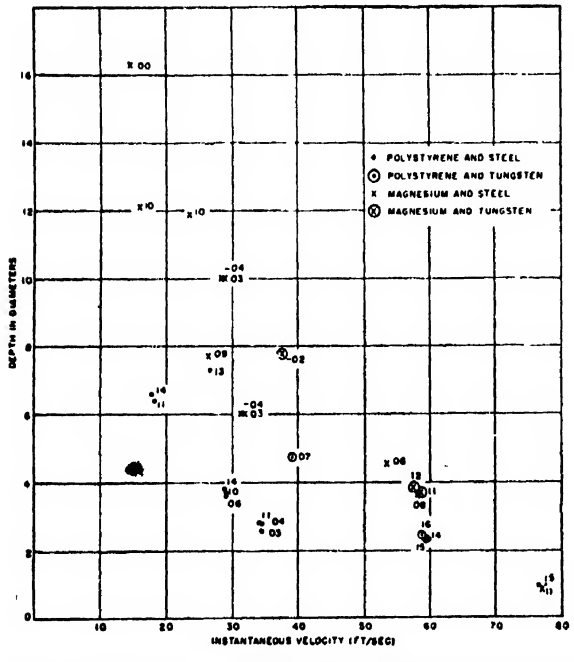


FIG. 2. Virtual-mass coefficient for vertical water-entry of spheres. Values (in percent) are plotted against the depth of the sphere and its velocity, at the instants for which the values were determined.

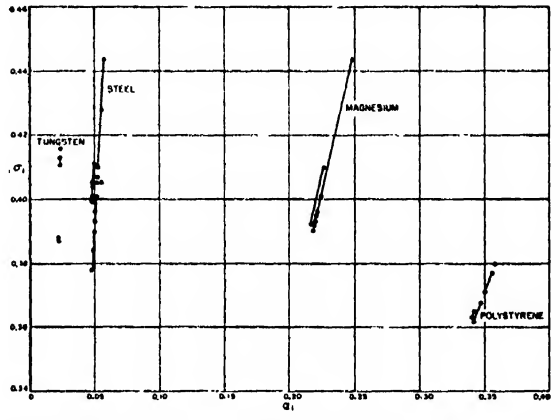


FIG. 3. Plot for the determination of virtual-mass coefficient on assumption that it has a single fixed value. The value required would be the negative of the slope of a straight line (not shown) passing through the center of several groups of points.

Before discussing the case of the water-entry cavity it is instructive to consider the steady-state cavity behind an obstacle in a water tunnel. One important difference between this case and the case of potential flow without a cavity is the absence of liquid immediately behind the body. There are still forces on the rear of the body due to water vapor and perhaps air in the cavity, but the "immediate" connection between the body and much of the water which it affects has disappeared. If the water throughout the tunnel were given a sudden increment of speed, water surrounding after-portions of the cavity would not be informed of the change of relative velocity between the water and the body for an appreciable time. The transverse diameter of any section of the cavity, therefore, depends not

so much on the present relative speed between body and water, as on the relative speed when that cross section of water was traversing the body's nose. (This effect is evident in the water-entry case shown in Fig. 1 where the diameters of the two cavities are nearly the same at the water surface because the speeds of entry were the same, and in spite of the greatly different speeds that existed later.) Because of this effect the virtual mass associated with steady-state cavity flow will be small.

Water which is set into motion by the body and is then without hydrodynamic contact with it, contributes to the drag but not to the virtual mass. It is evident from the geometry of the cavity that only a small amount of the water pushed out of the way by a body followed by a cavity, remains in hydrodynamic communication with the body, and this means simply that the virtual mass is small. While the cavity phase of water-entry is more difficult of treatment it is reasonable that it should have a virtual mass similar to that of the water-tunnel cavity.

CONCLUSION

The values reported here for virtual mass in the cavity phase of water-entry are small and show large percentage fluctuations. No systematic trends in the data and no reasons for the fluctuations have been found. However, it is not surprising that the small value of virtual mass should vary considerably because of the transient and variable character of the motion.

Because its magnitude is small and variable it appears improbable that virtual mass can help greatly in systematizing the data from water-entry experiments.

ACKNOWLEDGMENTS

The authors wish to express their indebtedness to members of the Hydroballistics Division, especially to Dr. J. Howard McMillen for stimulating discussions during the progress of this work, and to Mr. D. J. Milano who carried out the experimentation; also to Mr. Noel R. Corngold for his assistance in analyzing the data and to Professor Garrett Birkhoff for his many helpful discussions.

Plastic Flow of Platinum Wires

R. P. Carreker, Jr.

Research Laboratory, General Electric Company, Schenectady, New York

(Received June 16, 1950)

The strain-time behavior of annealed platinum wires at constant stress and temperature is reported for a wide range of experimental conditions, namely: 1550 to 78°K, 900 to 40,000 p.s.i., 0.001 to 0.1 strain, 10^{-1} to 10^{-6} min.$^{-1}$ strain rate.

The following relations were found to describe the results satisfactorily:

$$\sigma = A\dot{\epsilon}^n|_{\epsilon,T}; \quad \sigma = Be^{pT}|_{\epsilon,\dot{\epsilon}},$$
$$\epsilon = Ct^a|_{\sigma,T}; \quad \frac{\partial n}{\partial T} = \frac{\partial p}{\partial \log\dot{\epsilon}}\Big|_{\epsilon},$$
$$Q \sim \log\sigma|_{\epsilon};$$

where σ = stress, ϵ = strain, $\dot{\epsilon}$ = strain rate, T = absolute temperature, Q = activation energy for rate of deformation, t = time, and A, B, C, a, n, and p are constants.

INTRODUCTION

THERE is a paucity of good experimental data on the resistance to plastic deformation as a function of the amount, temperature, and rate of the deformation. Many experiments pertaining to deformation are reported in the literature, but surprisingly few data exist for which tests were performed on a single material with a definite prior history, under definitely stated and controlled conditions, over a sufficiently wide range of rate and temperature to be useful in formulating and testing theories of deformation.

This paper reports the results of a study of the strain-time behavior of a pure metal as a function of stress and temperature. The tests were of the short-time variety, lasting from a few seconds to several days.

MATERIAL

In view of the large number of identical specimens required, small wire specimens seemed appropriate. All specimens were taken from the same length of 99.98+ percent platinum[1] wire, which was cold-drawn through diamond dies from 0.020 to 0.015 in. diameter. After mounting in the testing equipment each specimen was annealed *in situ* for 10 min. at 1670°K by passing a predetermined electric current through the wire. This treatment produced a grain size of two to four grains per cross section.

TESTING EQUIPMENT

The testing apparatus was obtained from S. Dushman, of this Laboratory. A description of its original form has been published.[2] Extension of the wire specimen permits the movement of a grating parallel to a similar stationary grating, modulating a light beam

[1] Spectrographic analysis indicated traces (<0.01 percent) of Fe and Pd.

[2] Dushman, Dunbar, and Huthsteiner, J. App. Phys. 15, 108 (1944).

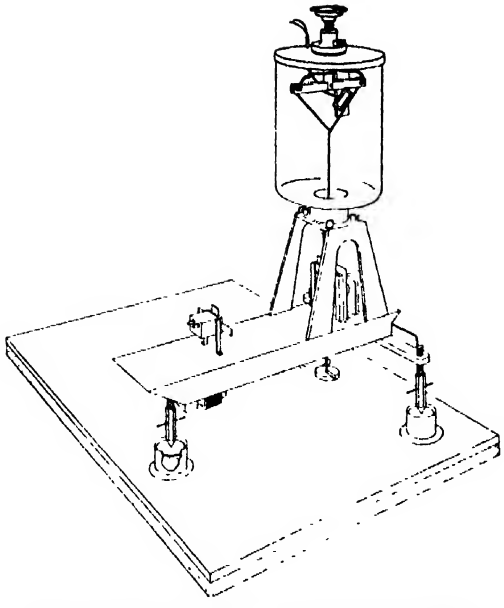


FIG. 1. Creep test equipment used above 1200°K.

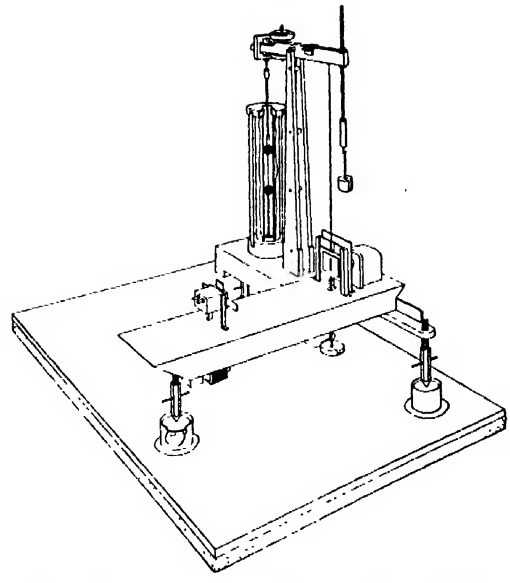


FIG. 2. Creep test equipment used between 300 and 1200°K.

striking a photo-cell. A change in light intensity from a minimum to a maximum corresponds to a grating displacement of 0.004 in. This system necessarily records only time-dependent strain.

Figure 1 shows the apparatus, as modified for use at temperatures above approximately 1200°K. In this arrangement the specimen is in the shape of a "V," supported at each end and loaded at the vertex. The specimen is heated internally by an electric current. Temperature of the specimen was indicated by a galvanometer which was activated by a photo-cell. The galvanometer was calibrated using an optical pyrometer as a secondary standard, correcting for emissivity and transmission effects. A black screen, not shown in Fig. 1, surrounded the glass test chamber, eliminating light from sources other than the specimen. Temperature control was achieved by bringing the specimen to the desired temperature, noting the reading of the indicating galvanometer, and manually adjusting the input voltage to correct any temperature change. Control within ± 10°K was attained.

Figure 2 shows the equipment as modified for use in the range 300 to 1200°K. A straight wire specimen was used, heated by a surrounding electric furnace that was controlled within ± 5°K by a Bristol potentiometer controller.

A liquid nitrogen bath was used to obtain the testing temperature of 78°K. In this case an insulated tank replaced the furnace of Fig. 2. Care was taken to maintain a constant bath level during each test. Temperatures were maintained within five degrees of the boiling point of liquid nitrogen, 78°K.

The choice of platinum as the specimen material eliminated the problem of oxidation at high temperatures. Pure dry nitrogen was passed through the furnace to increase the useful life of the stainless steel grips used with the straight wire specimens. The "V"-type specimens were tested in air.

CONSTANT STRESS

Two devices were used to maintain constant stress. The initial vertex angle of the "V" specimen was chosen to be 93°, so that constant stress was automatically maintained within 0.5 percent as the specimen elongated and the vertex angle became more acute.[3] The straight wire specimens were loaded through the bent lever shown in Fig. 2, whose dimensions were chosen to maintain constant stress within 0.2 percent as the specimen elongated.[4] Neither of these methods maintains constant stress after necking begins.

TESTING PROCEDURE

A. High Temperature "V"-Type Specimen

The specimen was mounted in the grips so that the distance between leg extremities was 20 cm. After annealing, the temperature was adjusted to the proper testing temperature by regulating the current flowing in the specimen. The load was placed on a weight pan hanging beneath the movable grating; a long, hooked wire (stainless steel 0.100 in. diameter) was then attached to the top of the movable grating and the hook lowered into the vertex of the "V." The displacement of the vertex was recorded as a function of time.

B. Straight Wire Specimens

The specimen was mounted between appropriate grips to give a gauge length of 3.0 in. The specimen and

[3] J. C. Fisher and R. P. Carreker, J. Metals 6, 178 (1949).
[4] Fullman, Carreker, and Fisher (to be published).

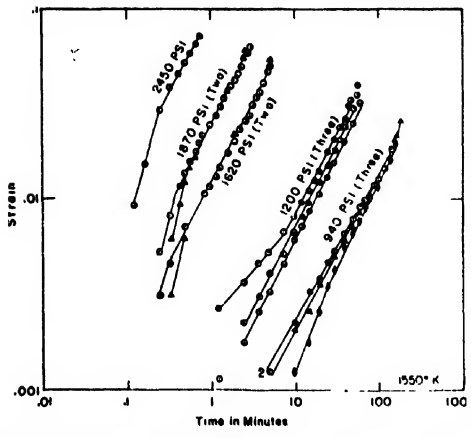


FIG. 3. Plastic strain of platinum as a function of time.

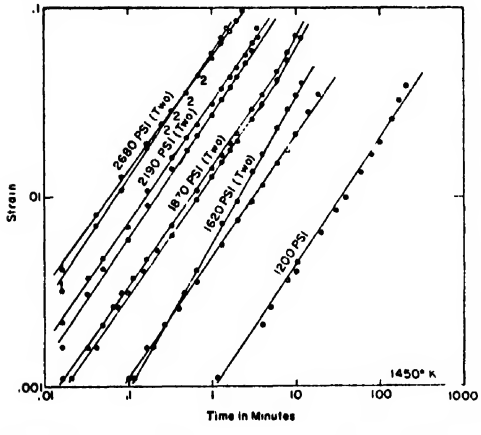


FIG. 4. Plastic strain of platinum as a function of time.

grips were lowered into the uniform temperature zone of the furnace, the lower grip securely fastened, and the upper grip attached to the constant stress lever. The position of the lever was such that no force was exerted on the specimen at this time. Electrical leads were fastened temporarily to the upper and lower grips for the annealing treatment. After removing the electrical leads from the specimen grips, the furnace (or tank) was brought from room temperature to the testing temperature. The selected load was then applied to the lever arm. Elongation of the specimen was recorded automatically as a function of time.

EXPERIMENTAL RESULTS

Several tests at different stresses were made at each of the following temperatures: 1550, 1450, 1350, 1250, 1100, 950, 850, 750, 650, 550, 450, 300, and 78°K. The experimental data are presented as plots of log time-dependent strain *vs.* log time in Figs. 3 to 15, each figure presenting the data obtained at a particular temperature.

Short-time points are absent in the data taken at 1550°K. In this series the specimens were first loaded at room temperature and the heating current then applied. The recorded movement was corrected for total thermal expansion. A short time elapsed before temperature equilibration was attained. Thus, this technique did not give satisfactory results during the first few seconds of the test. The 1550°K data have been included for their value at large strains. At all other temperatures the specimens were at temperature prior to loading.

In general, the log strain *vs.* log time plots are straight lines or approach straight lines at times in excess of one minute. The departures from linearity at short times are probably due to the difficulty in establishing the zero of time-dependent strain. This error is particularly noticeable at small strains and at times comparable with that required to apply the load to the specimen, approximately 1 to 2 sec. The error is negligible at strains greater than 0.01 and times greater than 1 min.

ANALYSIS OF EXPERIMENTAL RESULTS

The slopes of the log strain *vs.* log time plots, at strains not appreciably influenced by the zero error, depend strongly on temperature but are insensitive to

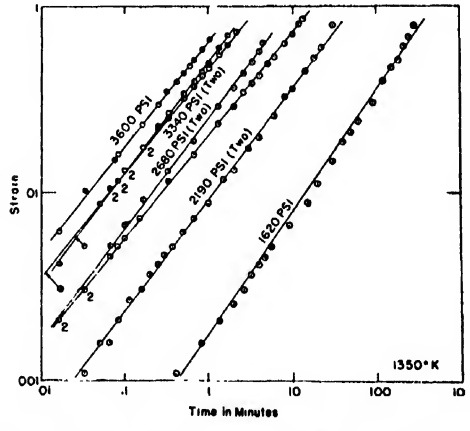


FIG. 5. Plastic strain of platinum as a function of time.

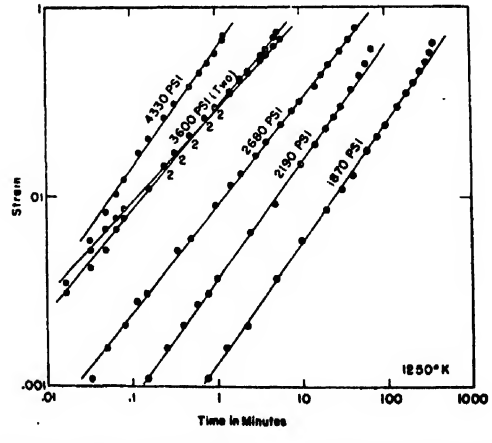


FIG. 6. Plastic strain of platinum as a function of time.

stress. These slopes have been plotted *vs.* temperature in Fig. 16. Figures 3 to 16 suggest the equation[5]

$$\epsilon = Ct^a,$$

where a is a number which apparently approaches 1.0 at the melting point and is small but positive at 0°K.

Consider the strain rates $\dot{\epsilon}$ of the various strain-time curves, at a particular value of the strain ϵ, as a function of the stress σ. At a strain of 0.01, the data from Figs. 3 to 15, when plotted as log stress *vs.* log strain rate, give the family of straight lines shown in Fig. 17, each line representing the results obtained at a particular temperature. Straight-line plots obtained in this manner imply that the relationship between stress and strain rate is of the form $\sigma = A\dot{\epsilon}^n|_{\epsilon, T}$. There is some scatter, but the trends are well established. Similar families were obtained by considering the results for strains equal to 0.005, 0.01, 0.02, and 0.03.

Consider the relationship between temperature and the stress necessary to produce a given strain rate at a given strain. The necessary data were obtained from Fig. 17 and plotted in Fig. 18. Log stress is linear in temperature, with some uncertainty below 450°K. Similar plots were obtained at strains of 0.005, 0.01, 0.02, and 0.03.

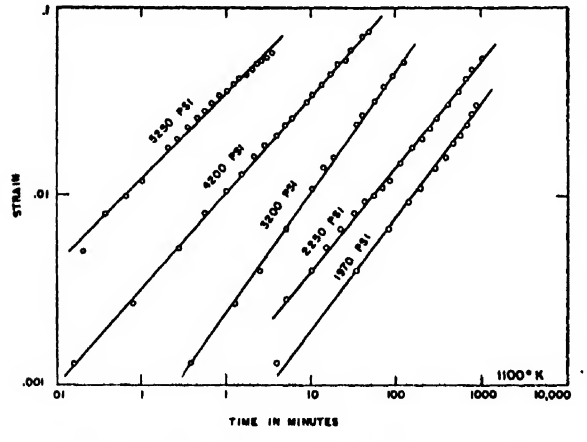


FIG. 7. Plastic strain of platinum as a function of time.

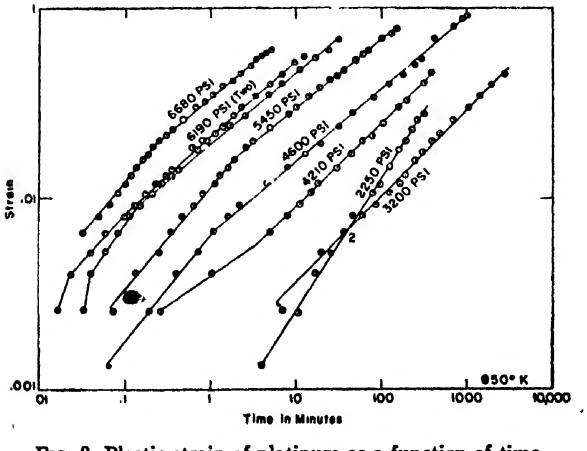


FIG. 8. Plastic strain of platinum as a function of time.

[5] The following notations are used throughout this analysis: t = time. σ = stress = applied force ÷ instantaneous area. ϵ = natural strain = $\ln(l/l_0)$. $\dot{\epsilon}$ = strain rate. T = absolute temperature. $e = 2.718 \cdots$. A, B, C, a, n, and p = constants.

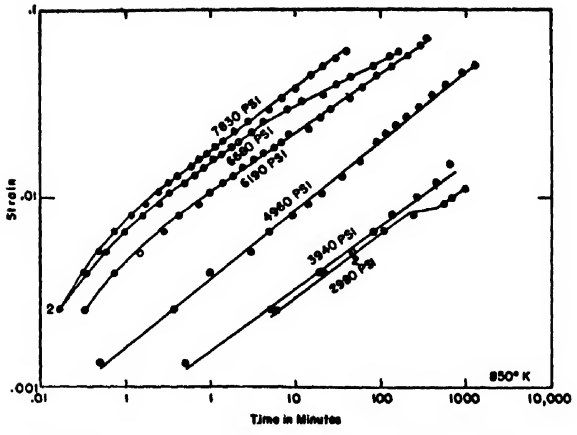


FIG. 9. Plastic strain of platinum as a function of time.

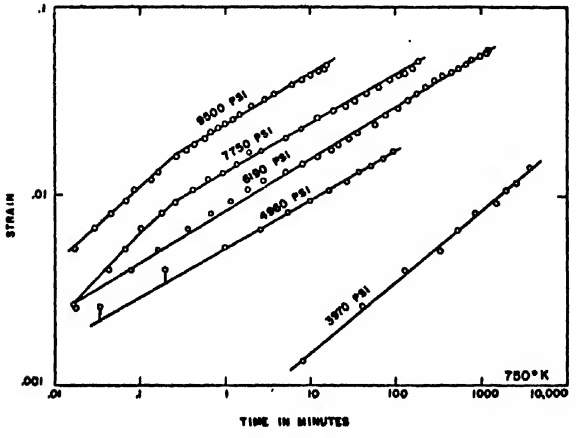


FIG. 10. Plastic strain of platinum as a function of time.

From consideration of Figs. 17 and 18, the following equations may be written:

$$\log\sigma = \log A + n\log\dot{\epsilon}|_{\epsilon, T} \tag{1}$$

and

$$\log\sigma = \log B + pT|_{\epsilon, \dot{\epsilon}}. \tag{2}$$

Previous research[6] has established the relation

$$\log\sigma = \log K + m\log\epsilon|_{\dot{\epsilon}, T}. \tag{3}$$

If it is assumed that stress is a point function of strain, strain rate, and temperature, one may write

$$d\log\sigma = \frac{\partial\log\sigma}{\partial\log\epsilon} d\log\epsilon + \frac{\partial\log\sigma}{\partial\log\dot{\epsilon}} d\log\dot{\epsilon} + \frac{\partial\log\sigma}{\partial T} dT, \tag{4}$$

or

$$d\log\sigma = m\,d\log\epsilon + n\,d\log\dot{\epsilon} + p\,dT. \tag{5}$$

[6] J. H. Hollomon, Metals Tech. 162, 268 (1945).

It is then possible to obtain analytically the general solution of Eq. (4).[7] Lubahn[8] has geometrically derived a general equation $\sigma=f(\dot{\epsilon}, \epsilon, T)$ using the relationships expressed in (1) and (3) and an expression for the activation energy for the deformation process. His equation may be written in the form

$$\sigma=CG^{T}(\dot{\epsilon}/\dot{\epsilon}_0)^{DT}e^{(E+E'T-FT\ \ln\dot{\epsilon}/\dot{\epsilon}_0)}, \qquad (6)$$

where C, G, $\dot{\epsilon}_0$, D, E, E', and F are constants. It should be noted that Eq. (6) includes the term $E'T$ which was omitted through a typographical error in the last equation of reference 8.[9] This equation is identical with that Leschen obtained analytically by the method of the exact differential and Eqs. (1) to (3).

The following relationships are obtained by differentiating the general equation (6). These relations offer alternative methods of determining the material constants from experimental data.

$$\frac{\partial^2 \log\sigma}{(\partial \log\dot{\epsilon})(\partial \log\epsilon)}=\frac{\partial m}{\partial \log\dot{\epsilon}}=\frac{\partial n}{\partial \log\epsilon}=-FT, \qquad (7)$$

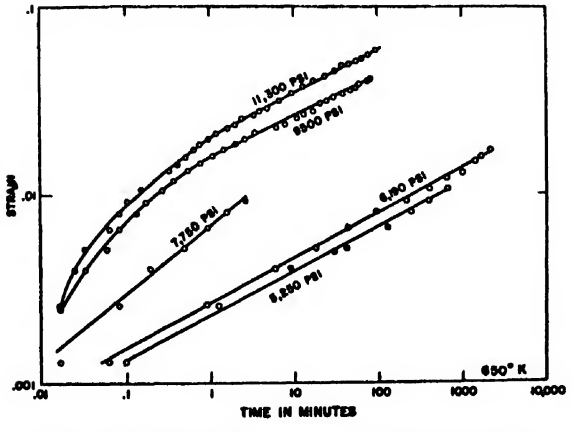


FIG. 11. Plastic strain of platinum as a function of time.

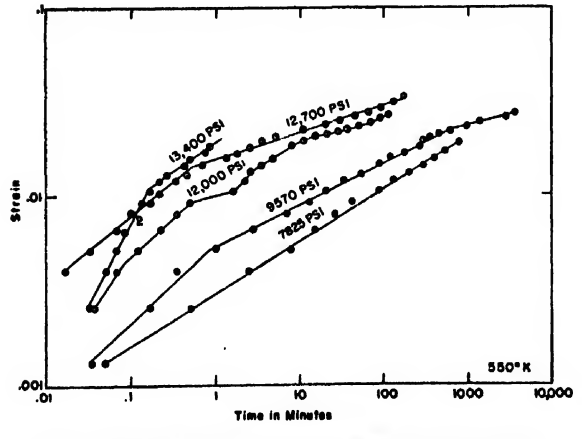


FIG. 12. Plastic strain of platinum as a function of time.

[7] J. G. Leschen (unpublished research).
[8] J. D. Lubahn, J. App. Mech. A-229 (September, 1947).
[9] C. Zener (private communication).

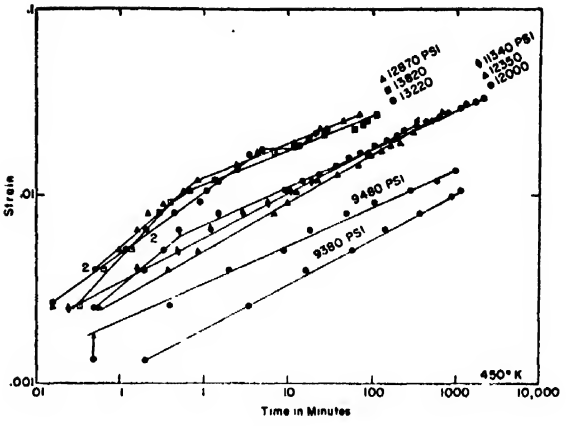


FIG. 13. Plastic strain of platinum as a function of time.

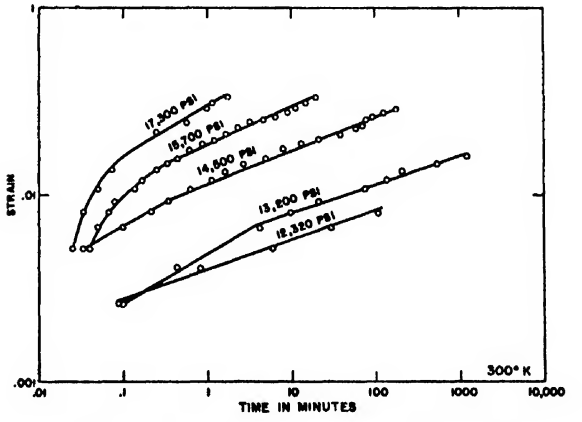


FIG. 14. Plastic strain of platinum as a function of time.

$$\frac{\partial^2 \log\sigma}{(\partial T)(\partial \log\dot{\epsilon})}=\frac{\partial n}{\partial T}=\frac{\partial p}{\partial \log\dot{\epsilon}}=D-F\log\epsilon, \qquad (8)$$

$$\frac{\partial^2 \log\sigma}{(\partial T)(\partial \log\epsilon)}=\frac{\partial m}{\partial T}=\frac{\partial p}{\partial \log\epsilon}$$

$$=E'-F\log\left(\frac{\dot{\epsilon}}{\dot{\epsilon}_0}\right), \qquad (9)$$

$$\frac{\partial^3 \log\sigma}{(\partial \log\epsilon)(\partial \log\dot{\epsilon})(\partial T)}=-F. \qquad (10)$$

Values of $\partial n/\partial T$ and $\partial p/\partial \log\epsilon$ may be obtained at constant values of strain. Values of n may be obtained from the slopes of the lines relating log stress to log strain rate at constant strain and temperature in Fig. 17. A plot of n *vs.* temperature is presented in Fig. 19. The points at 78°K are poorly defined due to the limited number of tests and small slopes at this temperature. Limits of 0.005 to 0.025 may be set with reasonable assurance. Values of p may be obtained from the slopes of the lines relating log stress to temperature at constant strain and strain rate in Fig. 18.

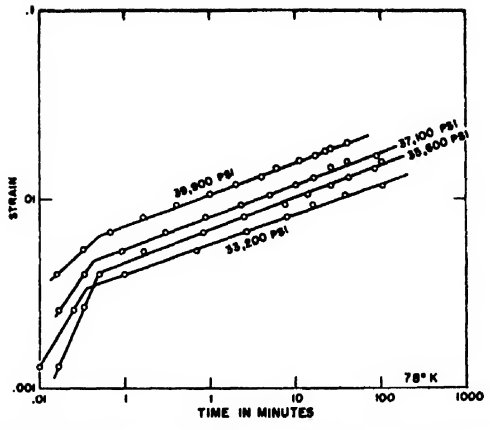


FIG. 15. Plastic strain of platinum as a function of time.

A plot of p *vs.* log strain rate is presented in Fig. 20. Equation (8) predicts that $\partial n/\partial T=\partial p/\partial \log\dot{\epsilon}|_{\epsilon}$. The following values for these quantities were obtained from Figs. 19 and 20:

$$\partial n/\partial T=1.04\times10^{-4}\ \text{deg.}^{-1}$$

and

$$\partial p/\partial \log\dot{\epsilon}=1.06\times10^{-4}\ \text{deg.}^{-1}.$$

Equations (7) to (9), together with Figs. 19 and 20, permit the quantity F to be taken as zero. Equation (6) then simplifies to the form

$$\sigma/\sigma_0=(\dot{\epsilon}/\dot{\epsilon}_0)^{DT}\epsilon^{(E+E'T)}. \qquad (11)$$

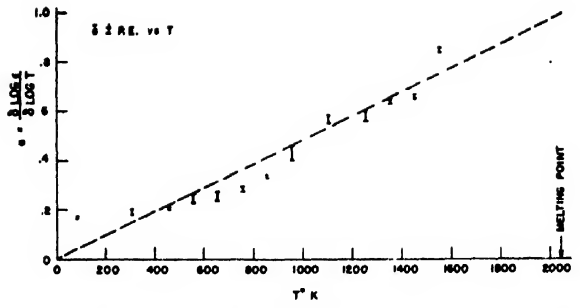


FIG. 16. Slopes of log strain-log time curves as a function of temperature (mean slope plus and minus probable error of mean).

In a creep test σ and T are constant and (11) can be reduced to $\epsilon=Ct^{\alpha}$, the form observed in the present experiments.

Many rate processes are dependent on temperature through a relationship of the form

$$\text{rate}\sim e^{-Q/RT},$$

where Q is the "activation energy" of the process per gram atom, $R=1.968$ cal. deg.$^{-1}$ mole^{-1} is the gas constant, and T is the absolute temperature. It may be instructive to consider deformation to be such a rate process. Then

$$d\epsilon/dt=\dot{\epsilon}=Ke^{-Q/RT}|_{\sigma,\epsilon}. \qquad (12)$$

Thus, a value of Q may be obtained from the slope of a

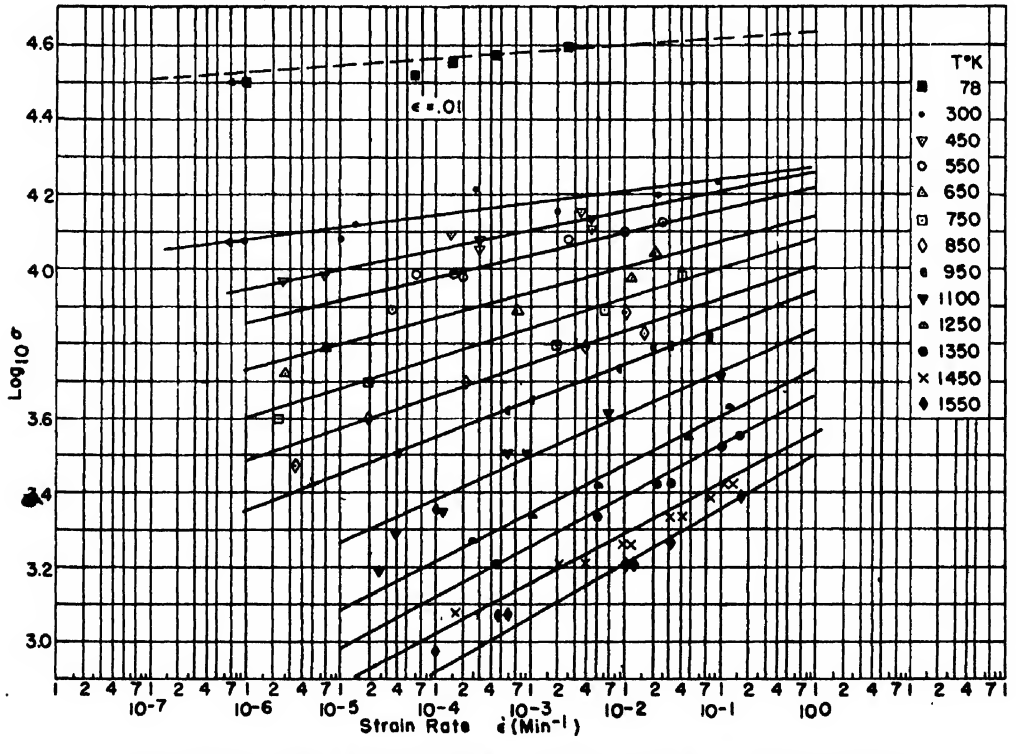


FIG. 17. Strain rate at a strain of 0.01 as a function of stress at several temperatures.

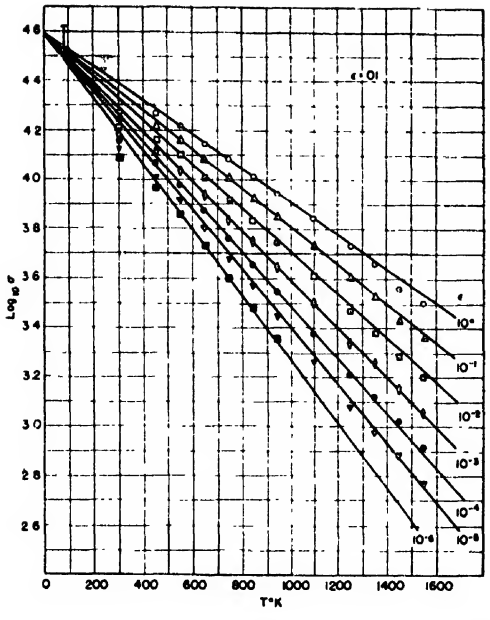


FIG. 18. Stress to produce several strain rates at a strain of 0.01 as a function of temperature.

plot of log$\dot{\epsilon}$ *vs.* $1/T$ at a particular value of stress and strain. The procedure may be repeated at other values of stress and strain to give the dependence of the activation energy on those variables. The data of Fig. 17 were cross plotted according to this procedure, resulting in Figs. 21 and 22. A similar procedure was followed considering strains of 0.005, 0.01, 0.02, and 0.03.

The form of the $Q=f(\sigma)$ relationship is of particular theoretical interest. Figure 22 shows that the decrease in Q with stress is not linear. The Becker[10]-Orowan[11] theory, in which attention is focused on the shear stress required to plastically deform an elementary region, predicts a linear relationship between $Q^{\frac{1}{2}}$ and σ. The data do not conform to this relationship. Nucleation theory, as applied to the formation of slip bands by Leschen, Carreker, and Hollomon,[12] predicts linearity between $Q^{\frac{1}{2}}$ and σ. The data do not confirm this prediction. The (empirical) equation (6) predicts Q to be linear in logσ, at constant strain.[13] Figure 23 shows the data plotted in this manner. The relationship is more nearly linear than the other functions of stress considered.

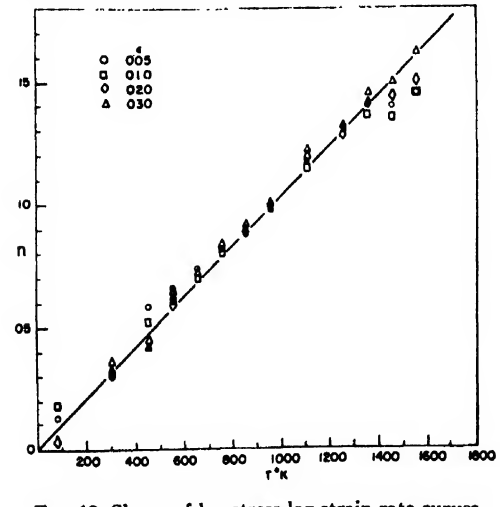


FIG. 19. Slopes of log stress-log strain rate curves as a function of temperature.

[10] R. Becker, Zeits. f. tech. Physik 7, 547 (1926).
[11] E. Orowan, Zeits. f. Physik 89, 605 (1934); 98, 382 (1936).

A CRITICAL APPRAISAL

There are definite undesirable features in the experimental techniques employed in this investigation. The most important of these is the narrow range of strain available for analysis. The upper limit of strain, 0.03, was dictated by the practical consideration of time per test at low strain rates and a desire to avoid strains near fracture, which occurred at 0.05 to 0.10. The lower value, a strain of 0.005, was limited by the accuracy with which the zero of time-dependent strain could be established. The zero of strain was established within 0.001.

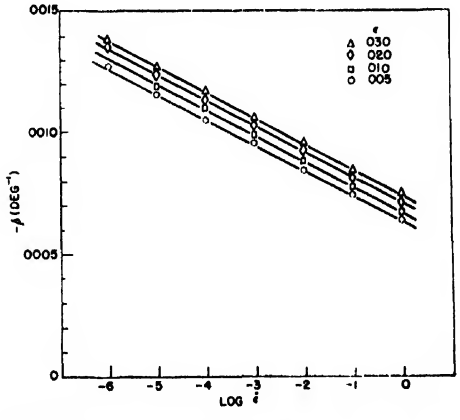


FIG. 20. Slopes of log stress-temperature curves as a function of strain rate.

The effective initial gauge lengths of the specimens were taken as 3.0 in. for the straight specimens and 5.57 in. for the "V" specimens, but these values are subject to some error. The wires were inserted in the grips and the lengths measured within 0.1 in. The grips would not be expected to influence the stress pattern in more than ten diameters (0.15 in.) of the specimen length. The "V" wires were heated by electric current and had cold ends where they contacted more massive pieces of metal. The distance affected by such end losses was estimated at less than 0.5 in. from each point of contact. Thus, the gauge lengths of the high tem-

[12] Leschen, Carreker, and Hollomon, Metals Tech. 15, T.P. 2476 (1948).
[13] J. H. Hollomon and J. D. Lubahn, Gen. Elec. Rev. (February and April, 1947).

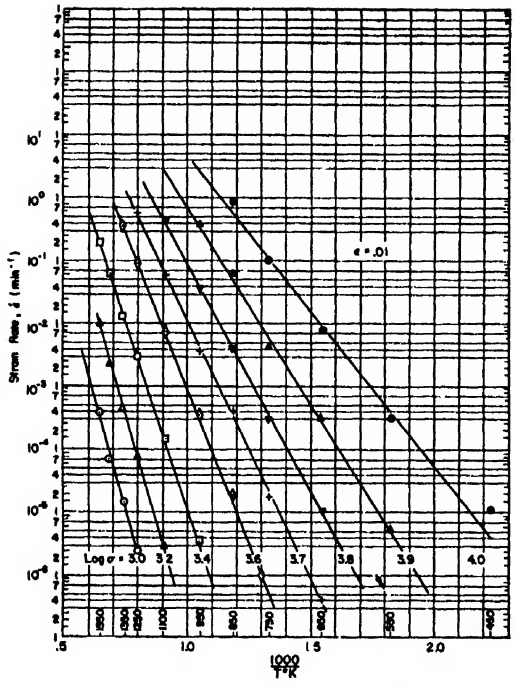


FIG. 21. Log strain rate at a strain of 0.01 *vs.* $1/T$ for several stresses.

perature "V"-type specimens were subject to a maximum error of 18 percent while the gauge lengths of the straight wire specimens were subject to a maximum error of eight percent. These errors might possibly lead to a maximum error of a factor of 1.2 in the strain rates. This error is negligible, considering the wide range of rates covered, and is certainly much less than the experimental scatter.

Although it is possible to measure and control temperature more accurately than was done in these experiments, it is felt that any errors in temperature are negligible with respect to the range of temperature explored.

It is believed that the experimental errors are insignificant with respect to the magnitude of the measured effects and that they do not obscure the basic description of the creep of platinum. The consistency and regularity of the data substantiate this belief.

The relations obtained from the present experiments were determined under a particular stress history, namely, constant stress. Tensile tests on the same material would be desirable as they would incorporate a different stress history and establish its effect, if any, upon these relations.

The assumption that σ is a point function of ϵ, $\dot{\epsilon}$, and T, which forms the basis of the foregoing empirical analysis, is not strictly correct. Dorn[14] has shown that the $\sigma-\epsilon$ relationship is influenced by temperature history when $\dot{\epsilon}$ is constant. In the present experiments σ and T were constant, during any one test, while ϵ and $\dot{\epsilon}$ were changing. The assumption appears to be valid under these conditions.

[14] J. E. Dorn *et al.*, Trans. A.I.M.E. 180, 205 (1949).

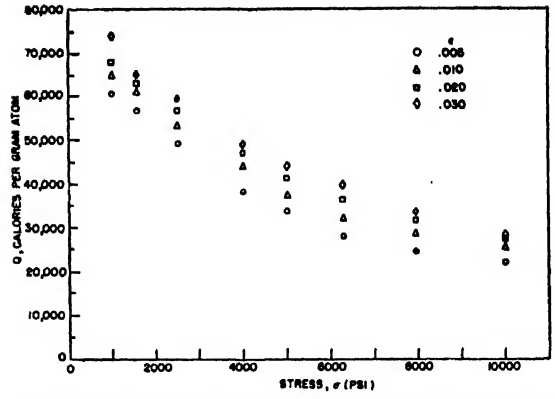


FIG. 22. Activation energy for plastic flow as a function of stress at several strains.

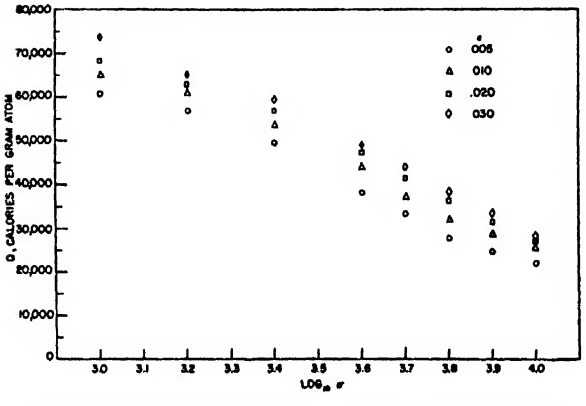


FIG. 23. Activation energy for plastic flow *vs.* logarithm of the stress at several strains.

It is interesting and important that the strain-time behavior of a metal subjected to such a wide range of experimental conditions can be described by simple relationships, even though these relationships are empirical. The present results indicate that useful information might be obtained by conducting similar experiments on other materials. Such experiments are in progress in this laboratory.

ACKNOWLEDGMENT

The writer wishes to acknowledge the assistance of Mrs. M. Wells in the experimental portion of this work; J. G. Leschen's contribution in developing the general equation relating stress, strain, strain rate, and temperature; and S. Dushman's contribution of equipment. The continued interest and helpful criticism of members of the Metallurgy Division of this Laboratory is sincerely appreciated.

The Effect of Magnetic Field on the Breakdown of Gases at Microwave Frequencies*

BENJAMIN LAX,† W. P. ALLIS, AND SANBORN C. BROWN
Research Laboratory of Electronics, Massachusetts Institute of Technology
(Received July 3, 1950)

The effect of magnetic field on the high frequency breakdown of gases has been studied. The presence of energy resonance and the modification of diffusion are shown experimentally and explained theoretically. An application is made of both the average electron theory and the Boltzmann theory, and the correspondence between these two theories is discussed.

THE breakdown of gases by high frequency electric fields in the presence of a constant magnetic field has been studied by Townsend and Gill[1] and by A. E. Brown.[2] It is the purpose of this paper to carry the analysis further, including the effect of the magnetic field on both the random diffusion of the electrons and their directed mobility.

Two approaches are available to such kinetic problems, and as there are advantages to each, both will be used. In Part I, the average electron theory will be given. In this method the orbit of a free electron in the assumed fields is computed first, and from this one computes the displacement and the energy gain in the time $\tau=t-t_0$ elapsed since a collision. These quantities are then averaged over the phase of the a.c. field at the time t_0 of the last collision, over the direction in space of the velocity after the collision, and over the free time up to the next collision. The result is the mean square displacement and energy gain of the average electron between collisions. One can then discuss an average electron from its initial low energy until it ionizes a gas atom or diffuses out of the tube. The condition for breakdown is that these two final achievements be equally probable. This method has the advantage that each step in the analysis has a direct physical meaning.

In Part II the Boltzmann transport equation is expanded in spherical harmonics in space, and in Fourier series in time. There results a differential equation for the distribution function which is integrated. Most of the properties of a discharge follow directly from a knowledge of the distribution function.

I. AVERAGE ELECTRON THEORY

Velocity between Collisions

Consider the motion of an electron between collisions under the influence of an electric field along the x-axis, $E=E_p\exp(j\omega t)$, and a constant magnetic field B along the z-axis. The equation of motion is then

$$\mathbf{F}=-e\mathbf{E}-ev\times\mathbf{B}=m\dot{\mathbf{v}}. \tag{1}$$

The solution of this equation is the sum of a general and a particular integral, which correspond to the superposition of a circular helical motion and a plane elliptical motion. For the helical motion whose axis is along the magnetic field, the velocity is

$$\left.\begin{aligned} v_{1x}&=(a+jb)\exp(j\omega_b\tau)\\ v_{1y}&=(b-ja)\exp(j\omega_b\tau)\\ v_{1z}&=c \end{aligned}\right\} \tag{2}$$

and oscillates at the cyclotron frequency $\omega_b=eB/m$. Because the helical motion contains the three arbitrary constants, the energy of this motion is constant and is given in electron-volts by:

$$u_1=\frac{mv_1^2}{2e}=\frac{m}{e}\frac{a^2+b^2+c^2}{2}. \tag{3}$$

For the elliptical motion, the velocity is

$$\left.\begin{aligned} v_{2x}&=\frac{eE_p}{m}\frac{j\omega}{\omega^2-\omega_b^2}\exp(j\omega t)\\ v_{2y}&=\frac{eE_p}{m}\frac{\omega_b}{\omega^2-\omega_b^2}\exp(j\omega t) \end{aligned}\right\} \tag{4}$$

and oscillates at the frequency of the applied field. The kinetic energy of this motion is uniquely determined by the magnitude and frequency of the applied field and is given by:

$$u_2=\frac{eE_p^2}{8m}\left[\frac{1}{(\omega-\omega_b)^2}+\frac{1}{(\omega+\omega_b)^2}+\frac{2\cos 2\omega t}{\omega_b^2-\omega^2}\right]. \tag{5}$$

The total energy $u=(m/2e)|(\mathbf{v}_1+\mathbf{v}_2)|^2$ will contain cross product terms u_{12} which are important but rather lengthy to write down. Their average value will be given later. The three constants, a, b, c, of the helical motion are determined by the velocity $\mathbf{v}_0=\mathbf{v}_1(\tau=0)+\mathbf{v}_2(t=t_0)$ immediately after a collision. As the time τ has been used in Eqs. (2) one has simply

$$\left.\begin{aligned} a&=v_{0x}-\frac{eE_p}{m}\frac{j\omega}{\omega^2-\omega_b^2}\exp(j\omega t_0)\\ b&=v_{0y}-\frac{eE_p}{m}\frac{\omega_b}{\omega^2-\omega_b^2}\exp(j\omega t_0)\\ c&=v_{0z}. \end{aligned}\right\} \tag{6}$$

* This work has been supported in part by the Signal Corps, the Air Materiel Command, and ONR.
† Now at Geophysical Directorate, USAF Cambridge Research Laboratories.
[1] J. S. Townsend and E. W. B. Gill, Phil. Mag. 26, 290 (1938).
[2] A. E. Brown, Phil. Mag. 29, 302 (1940).

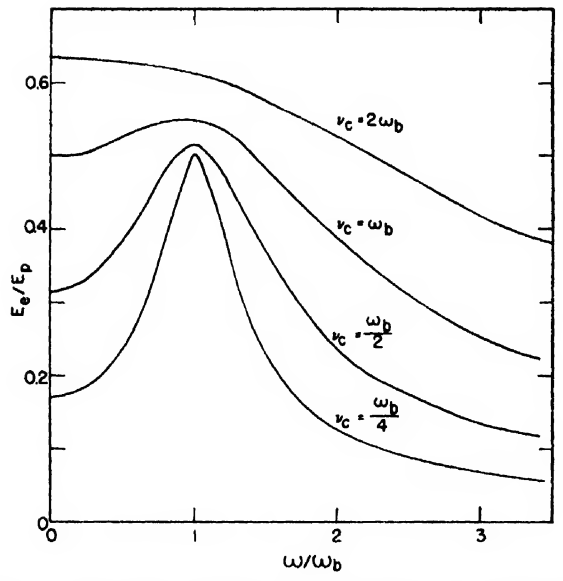


FIG. 1. The effective field E_e as a function of the frequency showing the resonance in the presence of a magnetic field.

It is noted that the elliptical motion exhibits a resonance at frequencies near the cyclotron frequency. Exactly at this frequency Eqs. (4) no longer hold and the solution corresponds to a spiral, but as collisions interrupt the motion it will not be necessary to use this singular solution.

Diffusion

From the velocities one obtains the displacements x_1, y_1, z_1, x_2, y_2 by integration. From these one can calculate the mean displacements $\langle x_1\rangle$, etc. but these all vanish on averaging over orientations of v_0, which is assumed isotropic, and over collision times t_0. In the average, an electron stays where it is in a high frequency discharge.

We are interested, however, in the mean square displacements $\langle x_1^2\rangle$, $\langle y_1^2\rangle$, $\langle z_1^2\rangle$ because these lead[3] to the diffusion coefficient. One finds the cross product terms such as $\langle x_1 y_1\rangle$ all vanish when averaged.

$$x_1^2+y_1^2=2\frac{a^2+b^2}{\omega_b^2}(1-\cos\omega_b\tau)$$
$$z_1^2=c^2\tau^2. \tag{7}$$

Averaging over orientations and times t_0, one finds that

$$\langle a^2\rangle=\langle b^2\rangle=\langle c^2\rangle=\frac{v_1^2}{3};\quad \langle ab\rangle=0. \tag{8}$$

The cross terms between the helical and elliptical motions also vanish but terms $\langle x_2^2\rangle$ and $\langle y_2^2\rangle$ do not vanish. However, these latter terms represent the mean square displacements due to mobility in the applied a.c. field and are not wanted in calculating the diffusion.

The average of a quantity X over the free times $\tau_c=1/\nu_c$ between collisions is, by definition,

$$\bar{X}=\int_0^\infty X\exp(-\nu_c\tau)\nu_c d\tau. \tag{9}$$

Applying this to the quantities in Eqs. (7) and defining the diffusion coefficients in terms of the mean square displacements, we obtain

$$D_{vv}=D_{xx}=\frac{\langle x_1^2\rangle}{2\tau_c}=\frac{v_1^2\nu_c}{3(\omega_b^2+\nu_c^2)}$$
$$D_{zz}=v_1^2/3\nu_c. \tag{10}$$

This definition of the diffusion tensor is of necessity symmetric. We shall see later that there are skew-symmetric terms which the random-walk definition cannot give. Diffusion along the z-axis is not altered by the magnetic field, but in the plane at right angles to the field it is reduced in the ratio $\nu_c^2/(\omega_b^2+\nu_c^2)$. For a given collision frequency, the diffusion coefficient is proportional to the energy of the electrons in their helical motion.

Energy Gain

The mean energy gain between collisions is best obtained by considering the power input to an electron $P=-e\mathbf{E}\cdot\mathbf{v}$. As the velocity v_{2x} is out of phase with the field, the corresponding power P_2 into this motion is zero in the average and only the power P_1 need be considered where

$$P_1=-eE_p\cos\omega t(a\cos\omega_b\tau-b\sin\omega_b\tau) \tag{11}$$

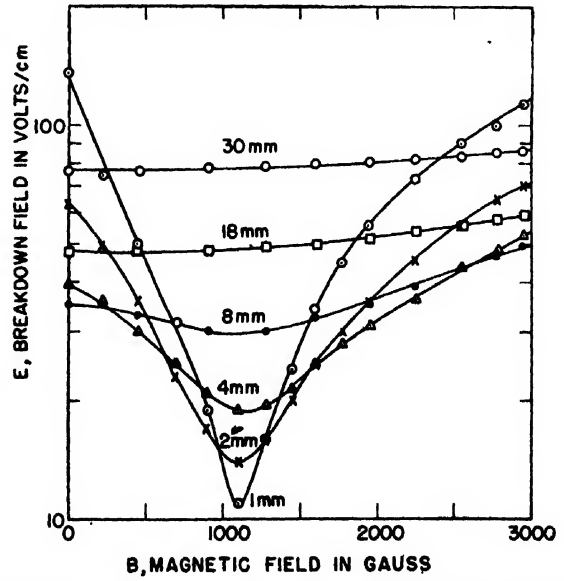


FIG. 2. Breakdown of helium in transverse electric and magnetic fields. (Diameter=7.32 cm and height=4.60 cm.)

[3] E. H. Kennard, *Kinetic Theory of Gases* (McGraw-Hill Book Company, New York, 1942), p. 286.

and the constants a and b are given by Eqs. (6). Averaging over orientations of the initial velocity the terms in v_0 drop out. Averaging over t_0 also we obtain

$$\bar{P}=\frac{e^2E_p^2}{4m}\left[\frac{\sin(\omega+\omega_b)\tau}{\omega+\omega_b}+\frac{\sin(\omega-\omega_b)\tau}{\omega-\omega_b}\right] \quad (12)$$

from which the average energy is obtained by integrating with respect to τ from 0 to τ.

$$\langle u_{12}\rangle=\frac{eE_p^2}{4m}\left[\frac{1-\cos(\omega+\omega_b)\tau}{(\omega+\omega_b)^2}+\frac{1-\cos(\omega-\omega_b)\tau}{(\omega-\omega_b)^2}\right]. \quad (13)$$

Averaging this quantity over collision times gives the mean energy gain between collisions

$$u_c=\frac{eE_p^2}{4m}\left[\frac{1}{(\omega+\omega_b)^2+\nu_c^2}+\frac{1}{(\omega-\omega_b)^2+\nu_c^2}\right]\equiv\frac{eE_e^2}{m\nu_c^2}. \quad (14)$$

This is a fundamental quantity in this theory. At low pressures ($\nu_c\to 0$) we see that it approaches twice the mean energy $\bar{u}_2$ of the elliptical motion of an electron, and in no case does it exceed this. At higher pressures, such that there are many collisions per oscillation, the energy $\bar{u}_2$ loses its meaning and the collision energy becomes $eE_p^2/2m\nu_c^2$. One can use Eq. (14) to define an effective field E_e which is the root-mean-square-field at high pressure. This concept is useful when the collision frequency ν_c is independent of velocity as this single function takes into account the effects of frequency and magnetic field on the energy.

At low pressures the effective field has a maximum at resonance with the cyclotron frequency as shown in Fig. 1.

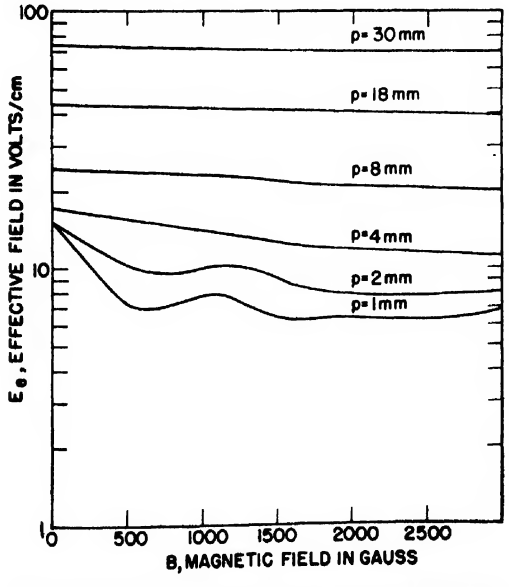


FIG. 3. Representation of experimental data of Fig. 2 in terms of the effective field.

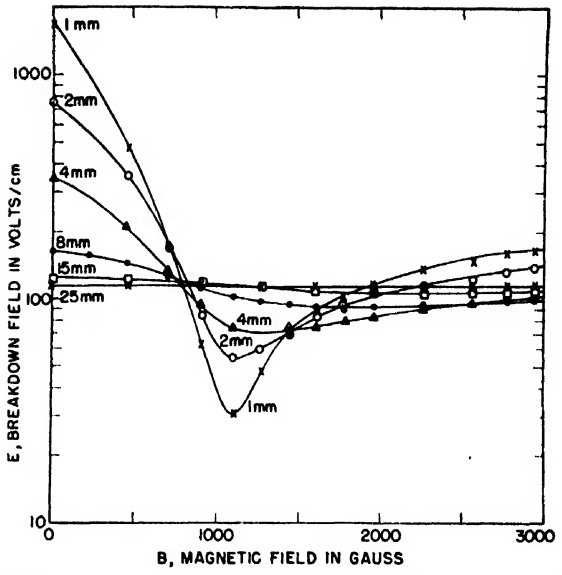


FIG. 4. Breakdown of helium in transverse electric and magnetic fields. (Diameter=7.32 cm and height=0.318 cm.)

Breakdown

The electrons produced by ionization have initially very little energy, but this increases by steps of u_c until the energy reaches a value u_i, at which ionization occurs. This is above the ionization potential V_i by an amount which we shall neglect. Excitations are disregarded in the following simple theory. The number N of free times to ionize is $N=u_i/u_c$ when ν_c is constant.

The electrons thus double their number by ionization every N collisions and unless some equally effective process exists which removes electrons their number will increase exponentially. In most cases diffusion to the walls of the discharge tube is the balancing process. In absence of the magnetic field, the random walk theory[4] gives the mean square distance $\Lambda^2=Nl^2/3$ reached in N free paths of mean square length l^2, so that if the average electron reaches the wall in a distance Λ the diffusion process will just balance ionization. This is the condition for breakdown and we can write it

$$\frac{u_c}{u_i}=\frac{\langle l^2\rangle}{3\Lambda^2}=\frac{\langle v^2\rangle}{3\Lambda^2\nu_c^2} \quad (15)$$

where Λ is now a length characteristic of the discharge tube and known as the diffusion length.[5]

If there is a magnetic field u_c will be altered according to Eq. (14). At the same time the random walk theory must be altered to take into account the curved paths between collisions. This may be done by appropriately decreasing l or increasing Λ. We shall adopt the latter and denote the new length by Λ_e. Its value will be given later.

[4] See reference 3, p. 271.
[5] M. A. Herlin and S. C. Brown, Phys. Rev. 74, 291 (1948).

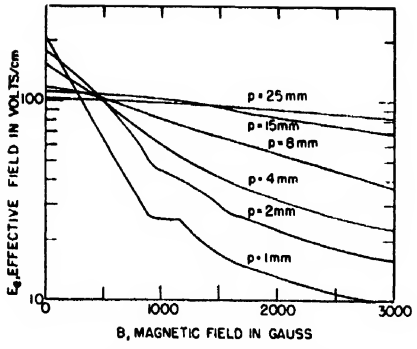


FIG. 5. Representation of experimental data of Fig. 4 in terms of the effective field.

When the mean free path is much smaller than Λ_e, the intercollision energy gain u_c is correspondingly smaller than the ionization potential. From Eq. (15) we see that breakdown should occur at the same effective field if the ratio of the mean free path to the effective diffusion length is the same, that is, the effective field for breakdown is a function of $p\Lambda_e$ only.

Combining Eqs. (14) and (15) we get

$$E_e^2=\frac{2\bar{u}u_i}{3\Lambda_e^2}. \tag{16}$$

Experimental data for breakdown in helium containing small admixtures of mercury vapor are shown in Fig. 2. The result of using the effective field with $\nu_e=2.37\times10^9 p$ as given by Brode's[6] collision probability measurements in place of the actual field but without taking the variation of Λ_e into account is shown in Fig. 3 for the same data. This shows that the resonance effect of the magnetic field and the high frequency are removed by using the effective field.

In order to test formula (10) for the diffusion coefficient, breakdown was studied in a flat cylindrical cavity whose length was very short compared to the radius. With the magnetic field placed transverse to the axis most of the diffusion has to take place perpendicular to the magnetic field and hence will show the full reduction. By Eq. (10) the mean square of the distance travelled by an electron is proportional to the diffusion coefficient D, and therefore the effective diffusion length Λ_e appropriate to infinite parallel plates is

$$\Lambda_e^2=\frac{\omega_b^2+\nu_e^2}{\nu_e^2}\Lambda^2. \tag{17}$$

The effect of a magnetic field is to make the dimensions of the cavity at right angles to the field appear larger to an electron. By Eq. (16) this should reduce the effective field for breakdown in the same proportion. Figure 4 shows a set of breakdown curves for helium in a cavity $2\frac{7}{8}$ in. in diameter and $\frac{1}{8}$ in. high. In Fig. 5 the effective field for the same data is plotted. In this figure the resonance peak has been removed as in Fig. 3, but now the curves are not horizontal because the magnetic field is increasing the effective cavity size.

However, Eq. (16) does not correspond with experiment except when used for comparative purposes. Equation (14) and the random walk theory give the number of free times for the average electron to ionize and reach the wall, respectively. But in a discharge it is the "faster-than-average" electron which ionizes and the "more-mobile-than-average" electron which leaves the tube and these are not the same electron. The mean free path method must therefore fail in predicting quantitative breakdown, and the failure should be worst when there are many collisions and therefore the greatest deviations from the mean.

II. BOLTZMANN THEORY

Spherical and Fourier Expansions

The Boltzmann transport equation is given by

$$C=\frac{\partial F}{\partial t}+\nabla\cdot(\mathbf{v}F)-\nabla_v\cdot\frac{e\mathbf{E}F}{m}-\nabla_v\cdot\frac{e}{m}\mathbf{v}\times\mathbf{B}F \tag{18}$$

where C is the production rate due to collisions per unit volume in phase space, ∇ and ∇_v are gradients in configuration and velocity space, respectively, and F is the distribution function. If F and C are expanded in spherical harmonics and substituted into the transport equation, we obtain two equations by equating the zero and first order terms

$$\begin{aligned}C_0&=\frac{\partial F_0}{\partial t}+\frac{v}{3}\nabla\cdot\mathbf{F}_1-\frac{e\mathbf{E}}{3mv^2}\cdot\frac{\partial(v^2\mathbf{F}_1)}{\partial v}\\ \mathbf{C}_1&=\frac{\partial\mathbf{F}_1}{\partial t}+v\nabla F_0-\frac{e}{m}\mathbf{E}\frac{\partial F_0}{\partial v}-\frac{e}{m}\mathbf{B}\times\mathbf{F}_1.\end{aligned} \tag{19}$$

We consider separately the elastic and inelastic collisions. Let C_{el} represent the elastic collision term. Morse, Allis and Lamar[7] have computed the zero- and first-order components of this term:

$$\begin{aligned}C_{0,el}&=\frac{m}{M}\frac{1}{v^2}\frac{\partial}{\partial v}\left(\frac{v^4F_0}{l}\right)\\ \mathbf{C}_{1,el}&=-\frac{v}{l}\mathbf{F}_1,\end{aligned} \tag{20}$$

where M is the mass of the gas molecule and l the mean free path of the electron.

The inelastic term C_{in} shows no angular dependence and hence, $C_{1,in}=0$ and only $C_{0,in}$ is retained. Equation

[6] R. B. Brode, Rev. Mod. Phys. 5, 243 (1933).

[7] Morse, Allis, and Lamar, Phys. Rev. 48, 412 (1935).

(19) then becomes

$$[0]\quad C_{0,in}+\frac{m}{M}\frac{1}{v^2}\frac{\partial}{\partial v}(\nu_c v^3 F_0)=\frac{\partial F_0}{\partial t}$$

$$+\frac{v}{3}\nabla\cdot\mathbf{F}_1-\frac{e}{3mv^2}\frac{\partial(v^2\mathbf{E}\cdot\mathbf{F}_1)}{\partial v} \quad (21)$$

$$[1]\quad 0=\nu_c\mathbf{F}_1+\frac{\partial\mathbf{F}_1}{\partial t}+v\nabla F_0-\frac{e}{m}\mathbf{E}\frac{\partial F_0}{\partial v}-\frac{e}{m}\mathbf{B}\times\mathbf{F}_1.$$

If the electric field is given by $\mathbf{E}=\mathbf{E}_p\exp(j\omega t)$, the Fourier expansions become

$$\mathbf{F}_1=\mathbf{F}_1^0+\mathbf{F}_1^1\exp(j\omega t)$$

and, with due caution in multiplying complex quantities,

$$\mathbf{E}\cdot\mathbf{F}_1=\tfrac{1}{2}\mathbf{E}_p\cdot(\mathbf{F}_1^1)_r+\mathbf{E}_p\cdot\mathbf{F}_1^0\exp(j\omega t),$$

where $(\mathbf{F}_1^1)_r$ is the real part of $\mathbf{F}_1^1$. These are now substituted into (21) and the Fourier components equated term by term. Only the constant term of the first equation is needed but the d.c. and a.c. terms of the second must be kept. In the steady state $\partial F_0/\partial t=0$.

$$[0,0]\quad C_{0,in}+\frac{m}{M}\frac{1}{v^2}\frac{\partial}{\partial v}(\nu_c v^3 F_0)-\frac{v}{3}\nabla\cdot\mathbf{F}_1^0$$

$$+\frac{e}{m}\frac{1}{3v^2}\frac{\partial}{\partial v}\left[\frac{v^2}{2}\mathbf{E}_p\cdot(\mathbf{F}_1^1)_r\right]=0 \quad (22)$$

$$[1,0]\quad \nu_c\mathbf{F}_1^0+v\nabla F_0-\frac{e}{m}\mathbf{B}\times\mathbf{F}_1^0=0$$

$$[1,1]\quad (\nu_c+j\omega)\mathbf{F}_1^1-\frac{e}{m}\mathbf{E}_p\frac{\partial F_0}{\partial v}-\frac{e}{m}\mathbf{B}\times\mathbf{F}_1^1=0.$$

These are the necessary equations for handling breakdown problems, which represent steady state conditions for the electrons. The above equations are applicable for any orientation of **E** and **B**. We shall only consider the cases when they are perpendicular or parallel to each other.

Diffusion Tensor

Integrating the [0,0] equations over velocity space in spherical coordinates, the second and fourth terms vanish at the limits. The first term gives the total production rate of electrons, $\nu_i n$, due to ionization. This is because excitations merely withdraw fast electrons to replace them by slow ones, whereas ionizations add an extra electron. The third term gives the divergence of a flow vector Γ

$$\Gamma=\int_0^\infty\frac{4\pi}{3}\mathbf{F}_1^0 v^3 dv \quad (23)$$

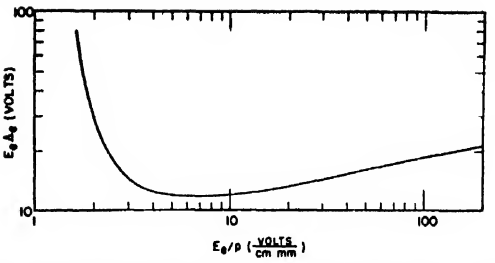


FIG. 6. Theoretical curve for the effective breakdown in helium.

so that

$$n\nu_i=\nabla\cdot\Gamma. \quad (24)$$

Solving Eq. [1,0] for $\mathbf{F}_1^0$ we find

$$\mathbf{F}_1^0=-\frac{v}{\nu_c}\frac{\nu_c^2+\omega_b\omega_b\cdot+\nu_c\omega_b\times}{\nu_c^2+\omega_b^2}\nabla F_0 \quad (25)$$

and this must be substituted in (23) and, assuming that F_0 can be written as a product $n(x,y,z)f_0(v)$ we find that Γ is proportional to ∇n but not necessarily in the same direction. Accordingly it is possible to define a diffusion coefficient but it must be a tensor

$$(D_{ij})=\int_0^\infty\frac{lv}{3(1+b^2)}$$

$$\times\begin{pmatrix}(1+b_x^2) & (b_xb_y-b_z) & (b_xb_z+b_y)\\ (b_xb_y+b_z) & (1+b_y^2) & (b_yb_z-b_x)\\ (b_xb_z-b_y) & (b_yb_z+b_x) & (1+b_z^2)\end{pmatrix}4\pi v^2 f_0 dv \quad (26)$$

$$\Gamma_i=-\textstyle\sum_j D_{ij}\dfrac{\partial n}{\partial j};\quad i,j=x,y,z$$

where $b_x=(\omega_b)_x/\nu_c$ and $b^2=b_x^2+b_y^2+b_z^2$. This expression reduces to the ordinary coefficient $D=\langle v^2\rangle/3\nu_c$ with no magnetic field when $b=0$. If the magnetic field is taken along the z-axis, it reduces to

$$D_{ij}=\int_0^\infty\frac{lv}{3(1+b^2)}\begin{pmatrix}1 & -b & 0\\ b & 1 & 0\\ 0 & 0 & 1+b^2\end{pmatrix}4\pi v^2 f_0 dv, \quad (27)$$

which is equivalent to Eq. (10) except that the present tensor has skew-symmetric terms which were not obtained by the random walk definition.

Substituting in (24) we obtain the diffusion equation

$$D_{xx}\frac{\partial^2 n}{\partial x^2}+D_{yy}\frac{\partial^2 n}{\partial y^2}+D_{zz}\frac{\partial^2 n}{\partial z^2}+\nu_i n=0 \quad (28)$$

which determines the spatial distribution of the electrons. One can use the normal diffusion coefficient D provided lengths are expanded at right angles to the magnetic field in the ratio $(1+b^2)^{\frac{1}{2}}$.

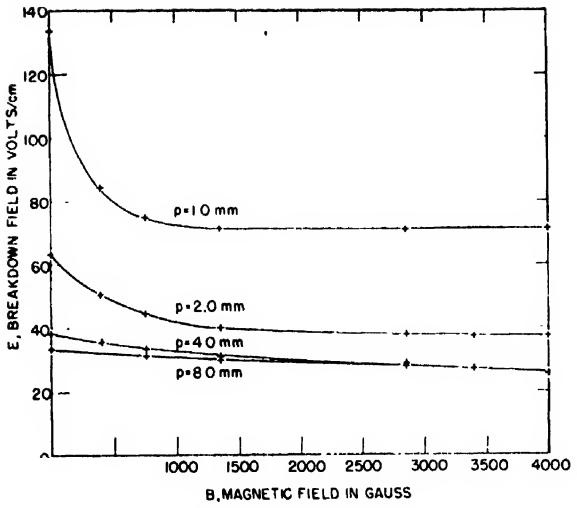


FIG. 7. Breakdown of helium in parallel electric and magnetic fields in a cylindrical cavity (diameter = 7.32 cm, height = 4.60 cm). The solid curves are theoretical and the points are experimental.

The solution of this equation depends on the boundary conditions. One must define an effective diffusion length Λ_e for the whole cavity which takes into account these expansions

$$\frac{1}{\Lambda_e^2}=\left(\frac{1}{\Lambda_x^2}+\frac{1}{\Lambda_y^2}\right)\frac{\nu_c^2}{\nu_c^2+\omega_b^2}+\frac{1}{\Lambda_z^2}. \tag{29}$$

The condition for the existence of a solution of (28) satisfying the boundary conditions is

$$D=\Lambda_e^2\nu_i \tag{30}$$

where D is the ordinary diffusion coefficient. The effect of the magnetic field is equivalent to expanding the cavity in the ratio $(\nu_c^2+\omega_b^2)^{\frac{1}{2}}\nu_c$ in all directions perpendicular to the magnetic field.

Differential Equation for F_0

The distribution function F_0 is obtained by eliminating $\mathbf{F}_1^0$ and $\mathbf{F}_1^1$ from Eqs. (22). One obtains in this way

$$C_{0,in}=\frac{v^2F_0}{3\nu_c\Lambda_e^2}-\frac{1}{v^2}\frac{\partial}{\partial v}\left[\frac{m}{M}\nu_c v^3F_0+\frac{e^2E_e^2}{3m^2\nu_c}v^2\frac{\partial F_0}{\partial v}\right] \tag{31}$$

where the effective field E_e is defined by (14) when E is at right angles to B. Should E make any other angle with B, the component at right angles is reduced by Eq. (14), and the square of this added to the mean square along B. Introducing the energy variable $u=mv^2/2e$ and the inelastic ratio $h=-C_{0,in}/\nu_eF_0$, Eq. (31) becomes

$$\frac{1}{(u)^{\frac{1}{2}}\nu_e}\frac{\partial}{\partial u}\left[\frac{2m}{M}\nu_c u^{\frac{1}{2}}F_0+\frac{2eE_e^2}{3m\nu_c}u^{\frac{1}{2}}\frac{\partial F_0}{\partial u}\right]=\left(h+\frac{2eu}{3m\nu_e^2\Lambda_e^2}\right)F_0. \tag{32}$$

The four terms of this equation are readily interpreted: the first, with $2m/M$, represents energy lost to recoil of the atoms, the second, with E_e^2, is the energy gain due to the field, h gives the loss of electrons through inelastic collisions, and the last term, in $1/\Lambda_e^2$, the loss through diffusion to the walls. It is noted that the magnetic field and the frequency are entirely contained in the effective quantities E_e and Λ_e. The pressure enters only in ν_c and indirectly through ν_c into E_e and Λ_e. The pressure cancels out of the first and third terms, and the only experimental parameters are thus E_e/p in the second term and $p\Lambda_e$ in the fourth. The other quantities are universal constants such as e/m and m/M and atomic constants such as the excitation and ionization cross sections included in h. It follows that all the breakdown data should plot on a single curve when $E_e\Lambda_e$ is plotted against E_e/p. This exact law is, however, of very limited applicability. The effective quantities E_e and Λ_e depend on ν_c, which is a function of the electron's energy. It is therefore in general impossible to make the effective values the same at all energies as the law requires. For helium and hydrogen the collision frequency ν_c is very nearly constant so the effective values are significant as shown in Fig. 6.

Because inelastic impacts set in discontinuously at the excitation potential u_x, it is necessary to solve Eq. (32) in two parts: Below u_x, $h=0$ and above u_x recoil and diffusion may generally be neglected. The solution of this equation in the completely analogous non-magnetic case has been carried out by MacDonald and Brown.[8]

The Boltzmann theory can be compared with the simple average electron theory in the case of helium

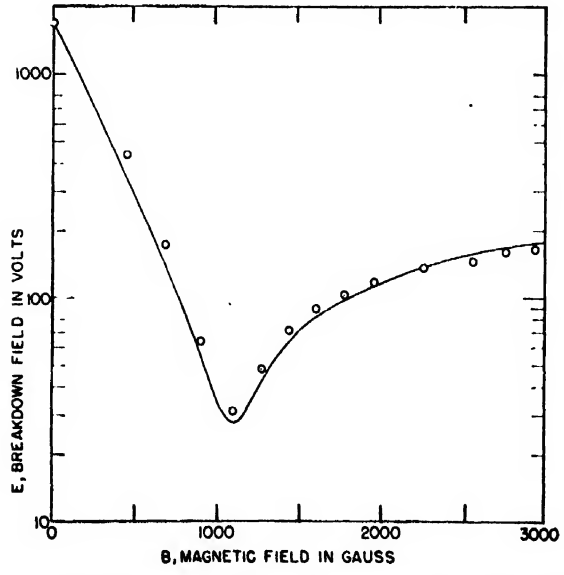


FIG. 8. Breakdown of helium at 1 mm pressure in a cylindrical cavity (diameter = 7.60 cm, height = 0.318 cm). Solid line is obtained from the Boltzmann theory and points are experimental.

[8] A. D. MacDonald and S. C. Brown, Phys. Rev. 75, 411 (1949).

containing traces of mercury so that all excitations result in ionizations. We must then neglect the recoil and excitation terms. The solution of the Eq. (32) is then expressible in terms of Bessel functions of order $\pm 1/4$. Setting the coefficients so that the function vanishes at u_i

$$F_0 = w^{-1/4}[J_{1/4}(w_i)J_{-1/4}(w) - J_{-1/4}(w_i)J_{1/4}(w)]$$

where

$$w = ju/\Lambda_e E_e.$$

The breakdown condition then reduces to

$$(1/4)!J_{1/4}(w_i) = 2w_i^{1/4}$$

or

$$u_i = 2.273\Lambda_e E_e.$$

This corresponds to using for the mean energy in (16)

$$\bar{u} = u_i/3.44.$$

III. NON-UNIFORM FIELDS

To study the effect of the magnetic field on diffusion alone the electric and magnetic fields are oriented in the same direction, and in order to reduce diffusion along the magnetic field it is necessary to perform the experiment in a cavity whose height is greater than the radius. In such a cavity the electric field may no longer be considered uniform and a correction to the computations must be made in a manner which has been shown by Herlin and Brown.[9] In the presence of the longitudinal magnetic field the equivalent diffusion length of the cylinder, from Eq. (29), is

$$\frac{1}{\Lambda_e^2} = \frac{\nu_c^2}{\nu_c^2 + \omega_b^2}\left(\frac{2.405}{R}\right)^2 + \left(\frac{\pi}{L}\right)^2.$$

Using this and the non-uniform field correction to the Boltzmann theory one obtains the agreement with experiment shown in Fig. 7. This result confirms the predicted effect of the magnetic field upon diffusion.

The breakdown measurements shown in Fig. 4 were made in a flat cavity with the magnetic field transverse to the axis. The effect of the latter is to require the solution of the diffusion equation in an elliptical cylinder whose diffusion length is then given by

$$\frac{1}{\Lambda_e^2} = \frac{\nu_c^2}{\nu_c^2 + \omega_b^2}\left[\left(\frac{\pi}{L}\right)^2 + \frac{1}{2}\left(\frac{2.405}{R}\right)^2\right] + \frac{1}{2}\left(\frac{2.405}{R}\right)^2.$$

Using this and the non-uniform field correction gives the theoretical curves shown in Fig. 8.

IV. EXPERIMENTAL APPARATUS AND PROCEDURE

The block diagram of the experimental equipment is shown in Fig. 9. The source of microwave energy is a

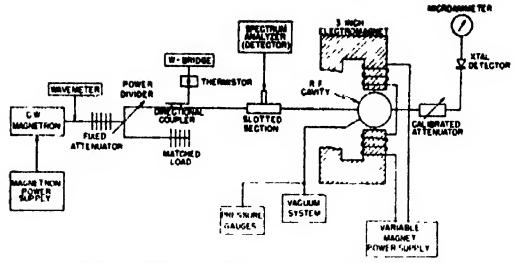


FIG. 9. Block diagram of the apparatus.

10-cm tunable c-w magnetron whose frequency is monitored by a calibrated wavemeter. The power fed into the cavity is varied by a balanced power divider and measured through a directional coupler by a bolometer, consisting of a thermistor and a balanced power bridge. The slotted section and detector are used to determine the electrical constants of the cavity.[10] The cavity, which is vacuum-tight, is loop coupled to the transmission line on either side. The output loop feeds into a variable calibrated attenuator, a crystal detector and a sensitive microammeter, which are adjusted to measure the power fed into the cavity at resonance. The magnetic field across the cavity is applied by an adjustable electromagnet. The magnetic field for a given gap size is calibrated as a function of the current through the coils by the use of a snatch coil and a ballistic galvanometer or by a compensated torque type fluxmeter. The pressure in the cavity during the experiment is measured by a McLeod gauge.

The breakdown experiment was performed at fixed values of pressure while the current through the coils and hence the magnetic field was set at different values. The power fed into the cavity was increased from nearly zero until the needle on the microammeter rose to a maximum and then suddenly dropped. This occurred at breakdown since the sudden increase in electron density detuned the cavity and produced a mismatch in the line. The setting of the calibrated attenuator and the maximum reading of the microammeter measured the power fed into the cavity. The breakdown field was then calculated from the measured constants and geometrical dimensions of the cavity.

V. CONCLUSIONS

The two principal effects of a magnetic field on high frequency breakdown of gases, the energy resonance with transverse fields and the reduction of diffusion, have been demonstrated experimentally and explained theoretically. The diffusion effect was shown to exist by itself when the electric and magnetic fields were parallel. The resonance phenomenon could not be separated because of the presence of diffusion at all times. Nevertheless this effect was brought into major

[9] M. A. Herlin and S. C. Brown, Phys. Rev. 74, 1650 (1948).

[10] S. C. Brown *et al.*, "Methods of measuring the properties of ionized gases at microwave frequencies," Technical Report No. 66, Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts.

prominence in breakdown by reducing the diffusion loss in a large cavity.

The development of the diffusion tensor in the presence of a magnetic field resulted in a more general diffusion equation. This led to the concept of the effective diffusion length, which together with the effective field, extended here to include the magnetic field, served to generalize the theory. The correspondence of the Boltzmann theory and the average electron theory was shown.

Quantum Limits of the Electrostatic Image Force Theory

R. G. Sachs and D. L. Dexter*
Department of Physics, University of Wisconsin, Madison, Wisconsin
(Received July 17, 1950)

The quantum mechanical problem of the interaction of a charged macroscopic body with a metal surface is treated by means of the variation method. The resulting expression for the interaction energy is the same as that obtained by the classical image force method except for three correction terms. These corrections are caused by the change in kinetic energy associated with the concentration of the electrons on the surface of the metal, the reduction in the electron-electron interaction energy in the metal resulting from the antisymmetrization of the wave function and the effect of the finite thickness of the surface charge. It is found that these terms are negligible except for distances between the charged body and metal surface such that surface structure effects are important.

I. INTRODUCTION

THE calculation of the interaction between a charged body and a much larger metal body is usually carried out by means of the classical image force theory. This treatment is certainly justified when the ratio of the distance between the bodies to atomic dimensions is sufficiently great. However, the method may be questioned when this ratio is of the same order of magnitude as the square root of the ratio of the total charge on the body to the electronic charge.[1] Such a condition may occur in experiments on the Schottky effect[2] and in the Van der Waals adsorption of molecules on a metal surface. The latter effect has been discussed by Lennard-Jones[3] from the point of view of the image force theory. A discussion of the limitations of this treatment due to the fact that the molecule must be treated quantum mechanically has been given by Bardeen.[4]

Bardeen has in addition treated the image force problem for an electron from the quantum mechanical point of view and has demonstrated that the classical formula for the force on an electron can be expected to hold asymptotically when the electron is at great distances from the metal. It is the purpose of this paper to determine the order of magnitude of the distance at which the classical formulas break down. In determining this distance, account will not be taken of the detailed surface structure of the metal, but the order of magnitude of the deviation from the classical formula due to purely quantum effects in the metal will be obtained.

This deviation is due to three effects. The first is the change in kinetic energy of the electrons that is associated with the increased charge density on the surface of the metal. The second effect is that due to the antisymmetrization of the wave function. This increases the average distance between electrons and therefore causes the electron-electron interaction to be smaller than that expected from classical considerations. Finally, in the determination of the self-interaction of the induced charge density account must be taken of the fact that one cannot assume an infinitely thin surface layer of charge in the quantum treatment since such an assumption would lead to an infinite kinetic energy of the electrons on the surface.

II. CLASSICAL TREATMENT

The classical theory of image forces will be briefly reviewed in this section from a point of view that may be easily compared with the quantum treatment. It will be assumed throughout the paper that a charged body (taken positive for the sake of definiteness), which gives rise to a potential $v(\mathbf{r})$ in the absence of other matter, interacts with the plane surface of a half-infinite metal body at distance R away. The origin of the coordinate system used is at the foot of the normal from the charged body to the metal surface. The positive z axis is taken as the normal into the metal and $\mathbf{s}$ is the two dimensional position vector on the surface of the metal.

In electrostatic theory it is shown that the potential $v(\mathbf{r})$ gives rise to a charge density, $\sigma_0(\mathbf{s})$, on the surface of the metal. The interaction energy, $W(R)$, is the sum of two terms. The first is the direct interaction between

* Research supported in part by ONR.
[1] Classical image force theory would not be expected to hold when the induced charge density, ρ', becomes equal to ρ_0, the charge density of free electrons in the metal. Since $\rho' \approx \sigma/a \approx Q/R^2 a$, where Q is the magnitude of the external charge, R its distance from the metal, and a is an atomic dimension, and since $\rho_0 \approx e/a^3$, ρ' becomes equal to ρ_0 when $Q/e \approx (R/a)^2$.
[2] E. Guth and C. J. Mullin, Phys. Rev. 59, 575 (1941).
[3] Lennard-Jones, Trans. Faraday Soc. 28, 334 (1932).
[4] J. Bardeen, Phys. Rev. 58, 727 (1940).

the charge density $\sigma_0(\mathbf{s})$ and the potential $v(\mathbf{r})$, and is given by

$$W_1(R)=\int \sigma_0(\mathbf{s})v(\mathbf{s})dS, \tag{1}$$

where the integration is carried out over the surface of the metal. The second term is due to the interaction of the charge density on the surface with itself:

$$W_2(R)=\frac{1}{2}\int\int \frac{\sigma_0(\mathbf{s})\sigma_0(\mathbf{s}')dSdS'}{|\mathbf{s}-\mathbf{s}'|}. \tag{2}$$

We can now make use of the fact that the total potential due to the charged body and the surface charge must vanish everywhere within the metal:

$$v(\mathbf{r})+\int \frac{\sigma_0(\mathbf{s}')dS'}{|\mathbf{r}-\mathbf{s}'|}=0, \tag{3}$$

for $Z>0$, from which it follows that

$$2\pi\sigma_0(\mathbf{s})=\frac{\partial v(\mathbf{r})}{\partial z}\bigg|_{z=0}.$$

Since the integral in Eq. (3) is a continuous function of s (although its derivative is not), we can set

$$v(\mathbf{s})=-\int \frac{\sigma_0(\mathbf{s}')dS'}{|\mathbf{s}-\mathbf{s}'|}.$$

Therefore

$$\int\int \frac{\sigma_0(\mathbf{s})\sigma_0(\mathbf{s}')dSdS'}{|\mathbf{s}-\mathbf{s}'|}=-\int v(\mathbf{s})\sigma_0(\mathbf{s})dS, \tag{3'}$$

and the total interaction energy is given by

$$W(R)=W_1(R)+W_2(R)=\frac{1}{2}\int \sigma_0(\mathbf{s})v(\mathbf{s})dS. \tag{4}$$

For future reference we wish to show that the energy $W(R)$ is a minimum with respect to variations in the charge density $\sigma(\mathbf{s})$. The variation of

$$W(R)=\int v(\mathbf{s})\sigma(\mathbf{s})dS+\frac{1}{2}\int\int \frac{\sigma(\mathbf{s})\sigma(\mathbf{s}')dSdS'}{|\mathbf{s}-\mathbf{s}'|}, \tag{5}$$

associated with the variation $\delta\sigma$ of σ is

$$\delta W(R)=\int v(\mathbf{s})\delta\sigma(\mathbf{s})dS+\frac{1}{2}\int\int \frac{\delta\sigma(\mathbf{s})\sigma(\mathbf{s}')dSdS'}{|\mathbf{s}-\mathbf{s}'|}+\frac{1}{2}\int\int \frac{\sigma(\mathbf{s})\delta\sigma(\mathbf{s}')dSdS'}{|\mathbf{s}-\mathbf{s}'|}.$$

Since the latter two integrals are equal, the condition that $\delta W(R)=0$ becomes

$$\int\left[v(\mathbf{s})+\int \frac{\sigma(\mathbf{s}')dS'}{|\mathbf{s}-\mathbf{s}'|}\right]\delta\sigma(\mathbf{s})dS=0,$$

and, taking into account the fact that the variations $\delta\sigma$ are arbitrary, this is equivalent to

$$v(\mathbf{s})+\int \frac{\sigma(\mathbf{s}')dS'}{|\mathbf{s}-\mathbf{s}'|}=0.$$

According to Eq. (3), this condition is satisfied by our choice of $\sigma_0(\mathbf{s})$. It follows that the energy given by Eq. (4) is a stationary value of Eq. (5); it also seems reasonable on physical grounds to assert that this stationary value is a minimum.

III. THE VARIATION METHOD

The wave equation for the metal electrons can be set up formally but, of course, it cannot be solved directly. The problem can, however, be treated by the variation method. In choosing a trial function, it will be assumed that the function leads to the classical surface charge density in very good approximation. All other details of the function will remain arbitrary. When this function is inserted into the variation integral, terms of the type (1) and (2) will occur. In addition there will be terms due to the change in kinetic energy, to the reduced interaction between electrons associated with the antisymmetric wave function, and to the non-zero thickness of the surface charge density. These latter terms turn out to be small compared to those given by Eqs. (1) and (2) as long as the distance between the charged body and metal surface is sufficiently great. Therefore, for such distances, the variation energy is essentially equal to (5). Since it has been shown in Section II that (4) is the minimum of Eq. (5) for variations in σ, it follows that, except for the small terms, this will also be the minimum of the energy with respect to variations in the wave function. It is, therefore, the proper energy associated with the wave equation as long as the correction terms may be neglected. Thus the quantum treatment leads directly to the same interaction as the classical image force theory.

In order to estimate the order of magnitude of the distance at which the image force theory may be expected to break down, it is necessary to seek that region in which the terms that have been neglected become of the same order of magnitude as the image force energy.

We consider as a variation function the normalized function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N)$ of the positions of the N electrons in the metal. The electronic charge density associated with this function is

$$\rho(\mathbf{r})=-e\sum_{j=1}^{N}\int\int\cdots\times\int|\Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_{j-1}, \mathbf{r}, \mathbf{r}_{j+1}, \cdots \mathbf{r}_N)|^2 d\tau^{N-1}, \tag{6}$$

where the $d\tau^{N-1}$ indicates that the integration is to be carried out over all the electrons except the j'th. e is the absolute value of the electronic charge. It is assumed that Ψ is chosen in such a way that $\rho=\rho_0+\rho'(\mathbf{r})$, where ρ_0, a constant, is the electronic charge density in the metal when the charged body is absent and $\rho'(\mathbf{r})$ is a small added volume density that arises in the presence of the charged body. It is also assumed that the charge distribution of the positive ions may be represented by a charge density $-\rho_0$, which is unchanged in the presence of the charged body. This assumption is certainly valid if $\rho'(\mathbf{r})$ varies slowly over the lattice spacing.

The general form of ΔE, the change in the energy of the system external charge plus metal, can be easily determined. There must be a term [see Eq. (1)]

$$\Delta E_V=\int \rho'(\mathbf{r})v(\mathbf{r})d\tau, \tag{1'}$$

corresponding to the interaction between the additional charge density $\rho'(\mathbf{r})$ and the potential $v(\mathbf{r})$. Also there is a term [see Eq. (2)]

$$\Delta E_R=\frac{1}{2}\int\int \frac{\rho'(\mathbf{r})\rho'(\mathbf{r}')d\tau d\tau'}{|\mathbf{r}-\mathbf{r}'|}, \tag{2'}$$

for the self-interaction of the additional charge density. The change ΔT in the kinetic energy of the electrons associated with a change in the wave function will have the form of a volume integral of some function of ρ over the metal, minus a volume integral of the same function of ρ_0 over the metal as long as changes in ρ are small within an electron wave-length. Expanding these functions in power series in ρ' and integrating, it is clear that the leading term will be of the form $\int\rho'^2d\tau$ since the terms involving only ρ_0 will cancel and the volume integral of ρ' is zero. Finally there must be a correction term ΔE_A to Eq. (2'), taking account of the reduced interaction between the metal electrons due to correlations of pairs of electrons. By the same argument used in connection with ΔT, ΔE_A will have a leading term $\int\rho'^2d\tau$. Thus

$$\Delta E=\int \rho'(\mathbf{r})v(\mathbf{r})d\tau + \frac{1}{2}\int\int \frac{\rho'(\mathbf{r})\rho'(\mathbf{r}')d\tau d\tau'}{|\mathbf{r}-\mathbf{r}'|}+\kappa\int\rho'(\mathbf{r})^2d\tau, \tag{7}$$

where κ is a constant depending on the detailed quantum properties of the metal electrons.

To estimate the order of magnitude of κ, we apply the Fermi-Thomas-Dirac statistical model to the metal electrons. According to this model, if the electron density is $n=-\rho/e$, the total kinetic energy of the electrons in the metal is

$$T=\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}\left(\frac{3h^2}{10m}\right)\int n^{5/3}d\tau,$$

where the integration is carried out over the volume of the metal. Therefore, for a small change, $n'=-\rho'/e$, in n the change in T is

$$\Delta T=\frac{h^2}{2m}\left(\frac{3n_0}{8\pi}\right)^{\frac{1}{3}}\int n'd\tau+\frac{h^2}{6mn_0}\left(\frac{3n_0}{8\pi}\right)^{\frac{1}{3}}\int n'^2d\tau.$$

The first term vanishes because of the condition for the conservation of charge and we are left with

$$\Delta T=\frac{(2\pi)^2}{6}\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}\frac{h^2}{me^2}n_0^{-\frac{1}{3}}\int\rho'(\mathbf{r})^2d\tau. \tag{8}$$

ΔE_A can be estimated in a similar way. The change in the electron-electron interaction due to the antisymmetrization of plane electron waves is

$$-\frac{3}{2}\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}e^2\int n^{4/3}d\tau,$$

and the change in this quantity associated with a change, n', in n can be obtained just as before. The result is

$$\Delta E_A=-\frac{1}{3}\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}n_0^{-\frac{1}{3}}\int\rho'(\mathbf{r})^2d\tau; \tag{9}$$

so

$$\kappa\approx-\frac{1}{3}\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}n_0^{-\frac{1}{3}}+\frac{(2\pi)^2}{6}\left(\frac{3}{8\pi}\right)^{\frac{1}{3}}\frac{h^2}{me^2}n_0^{-\frac{1}{3}}. \tag{10}$$

κ is of the order of the square of the inter-electronic distance for the free electrons and is a positive quantity; letting β denote the distance into the metal within which ρ' differs appreciably from zero, it is clear that in order to use the Fermi-Thomas approximation β must be larger than the free electron wave-length, or $\beta\gtrsim\kappa^{\frac{1}{2}}$. If β is the distance over which ρ' is not ≈ 0, ρ' in this region is of the order $\sigma(\mathbf{s})/\beta$; then the third term in Eq. (7) is of the order βR^{-1} times the first or second term, and for R large enough it is a small correction. Thus for large R, ρ' will be closely equal to the classical value $\sigma_0(\mathbf{s})\delta(z)$ where $\delta(z)$ is the Dirac delta function. But because of the third term in Eq. (7) we expect ρ' to have a small but non-zero extension in the z direction. The thickness of this region will be estimated below.

In order to evaluate ΔE, let us approximate $\rho'(\mathbf{r})$ by

$$\begin{aligned}\rho'(\mathbf{r})&=\beta^{-1}\sigma(\mathbf{s}) \quad \text{for} \quad z<\beta,\\ &\approx 0 \quad \text{for} \quad z>\beta,\end{aligned}$$

where $\sigma(\mathbf{s})$ is a surface charge density. Then the variation $\delta\Delta E$ in the energy ΔE associated with the variation

$\delta\beta$ in β and $\delta\sigma(\mathbf{s})$ in $\sigma(\mathbf{s})$ is

$$\delta\Delta E=\Big\{-\frac{1}{\beta^2}\int\int_0^\beta v(\mathbf{r})\sigma(\mathbf{s})d\tau+\frac{1}{\beta}\int v(\mathbf{s},\beta)\sigma(\mathbf{s})dS$$
$$-\frac{1}{\beta^3}\int_0^\beta\int_0^\beta\int\int\frac{\sigma(\mathbf{s})\sigma(\mathbf{s}')dSdS'dzdz'}{|\mathbf{r}-\mathbf{r}'|}$$
$$+\frac{1}{\beta^2}\int_0^\beta\int\int\frac{\sigma(\mathbf{s})\sigma(\mathbf{s}')dz'dSdS'}{((\mathbf{s}-\mathbf{s}')^2+(\beta-z')^2)^{\frac{1}{2}}}$$
$$-\frac{\kappa}{\beta^2}\int\sigma^2(\mathbf{s})dS\Big\}\delta\beta+\int\Big\{\frac{1}{\beta}\int_0^\beta v(\mathbf{r})dz$$
$$+\frac{1}{\beta^2}\int_0^\beta\int_0^\beta\int\frac{\sigma(\mathbf{s})dzdz'dS'}{|\mathbf{r}-\mathbf{r}'|}+\frac{2\kappa}{\beta}\sigma(\mathbf{s})\Big\}\delta\sigma(\mathbf{s})dS.$$

This leads to the following two conditions to be satisfied for minimum energy:

$$\int\int_0^\beta v(\mathbf{r})\sigma(\mathbf{s})d\tau-\beta\int v(\mathbf{s},\beta)\sigma(\mathbf{s})dS$$
$$+\frac{1}{\beta}\int_0^\beta\int_0^\beta\int\int\frac{\sigma(\mathbf{s})\sigma(\mathbf{s}')d\tau d\tau'}{|\mathbf{r}-\mathbf{r}'|}$$
$$-\int\int\int_0^\beta\frac{\sigma(\mathbf{s})\sigma(\mathbf{s}')dSdS'dz'}{((\mathbf{s}-\mathbf{s}')^2+(\beta-z')^2)^{\frac{1}{2}}}$$
$$+\kappa\int\sigma^2(\mathbf{s})dS=0, \quad (11A)$$

$$\int_0^\beta v(\mathbf{r})dz+\frac{1}{\beta}\int_0^\beta\int_0^\beta\int\frac{\sigma(\mathbf{s}')dzdz'dS'}{|\mathbf{r}-\mathbf{r}'|}$$
$$+2\kappa\sigma(\mathbf{s})=0. \quad (11B)$$

To zero order in β Eq. (11B) reads

$$v(\mathbf{s})+\int\frac{\sigma(\mathbf{s}')dS'}{|\mathbf{s}-\mathbf{s}'|}=0,$$

and since this is the condition satisfied by the classical surface charge density, $\sigma(\mathbf{s})$ may be written as $\sigma(\mathbf{s})=\sigma_0(\mathbf{s})+\sigma_1(\mathbf{s})$ where $\sigma_0(\mathbf{s})$ is the classical density and $\sigma_1(\mathbf{s})$ is a first order correction term.

For convenience let us define the quantity

$$\theta(\mathbf{r})\equiv v(\mathbf{r})+\frac{1}{\beta}\int_0^\beta dz'\int\frac{\sigma(\mathbf{s}')dS'}{|\mathbf{r}-\mathbf{r}'|}, \quad z\gtrless 0.$$

This is the total potential within the metal, and to a zero order approximation $\theta(\mathbf{r})$ is clearly zero. Using this definition (11A) and (11B) become

$$\int\sigma(\mathbf{s})\theta(\mathbf{s},\beta)dS=-\kappa/\beta\int\sigma^2(\mathbf{s})dS, \quad (12A)$$

$$\int_0^\beta\theta(\mathbf{r})dz=-2\kappa\sigma(\mathbf{s}). \quad (12B)$$

Now according to Eq. (3) θ may be written

$$\theta(\mathbf{r})=v(\mathbf{r})-\frac{1}{\beta}\int_0^z dz'v(\mathbf{s},z-z')$$
$$-\frac{1}{\beta}\int_z^\beta dz'v(\mathbf{s},z'-z)+\frac{1}{\beta}\int_0^\beta\int\frac{dz'\sigma_1(\mathbf{s}')dS'}{|\mathbf{r}-\mathbf{r}'|}.$$

Since $v(\mathbf{r})$ is a well-behaved function of z, we may expand in a Taylor's series about $z=z'$ and obtain

$$\theta(\mathbf{r})=-2\pi\sigma_0(\mathbf{s})\Big[z^2-2z\beta+\frac{\beta^2}{2}\Big]$$
$$+\frac{1}{\beta}\int_0^\beta dz'\int\frac{\sigma_1(\mathbf{s}')dS'}{|\mathbf{r}-\mathbf{r}'|}. \quad (13)$$

Introducing this expression into (12B) results in

$$\int\frac{\sigma_1(\mathbf{s}')dS'}{|\mathbf{s}-\mathbf{s}'|}=-\frac{2\kappa}{\beta}\sigma_0(\mathbf{s})-\frac{\pi\beta}{3}\sigma_0(\mathbf{s}). \quad (14)$$

Putting (13) and (14) in (12A) and integrating, we obtain the thickness of the surface layer

$$\beta=(3/2\pi)^{\frac{1}{2}}\kappa^{\frac{1}{2}}.$$

Now the expression for ΔE (Eq. 7) can be readily evaluated. Substituting (11) in (7) we obtain

$$\Delta E=\frac{1}{2\beta}\int_0^\beta dz\int\sigma(\mathbf{s})v(\mathbf{r})dS$$
$$+\frac{1}{2\beta}\int_0^\beta dz\int\sigma(\mathbf{s})\theta(\mathbf{r})dS+\frac{\kappa}{\beta}\int\sigma^2(\mathbf{s})dS$$
$$=W(R)+\frac{1}{2}\int\sigma_1(\mathbf{s})v(\mathbf{s})dS+\frac{\pi\beta}{2}\int\sigma^2(\mathbf{s})dS,$$

and using (14), there results

$$\Delta_1E\equiv\Delta E-W(R)=(8\pi/3)^{\frac{1}{2}}\kappa^{\frac{1}{2}}\int\sigma^2(\mathbf{s})dS. \quad (15)$$

Since $\int\sigma^2(\mathbf{s})dS$ is of the order $R^{-1}|W(R)|$, this equation shows that for R large compared with the spacing between the free metal electrons the quantum mechanical correction Δ_1E to the classical image energy $W(R)$ is small.

In the case of copper, n_0 is about one free electron per 100 $(\hbar^2/me^2)^3$, and the plane wave estimate of κ is 3.7 $(\hbar^2/me^2)^2$. Thus β becomes 1.3 $\hbar^2/me^2$ or $\beta \approx a$ where a is the lattice spacing. It is apparent that this is not consistent with the assumption we made when we replaced the positive ion charge density by $-\rho_0$, i.e., $\beta \gg a$. But since the region of high electron density coincides with a surface layer in which the positive ion density is larger than $-\rho_0/e$, it appears that we have underestimated the absolute magnitude of the (negative) interaction energy between the electrons and the positive ions. That is, the energy $\Delta_1 E$ which we have calculated appears to be too large. Thus it would seem that by taking into account the discrete character of the positive ions, energy could be gained by further compressing the charge density near the surface. However, it must be recognized that the electrons with which we form the surface density are taken from the top of the Fermi lake, or, since we are dealing with a metal, from the middle of an energy band. The electrons in the lower part of the band concentrate around the positive ions and provide enough shielding so that the discrete character of the ions is pretty well washed out in its influence on the wave function of a mid-band electron. Thus our estimate of β and $\Delta_1 E$ should be reasonable as long as just a few electrons near the top of the Fermi lake are influenced by the presence of the external field.

Since our calculated $\Delta_1 E$ becomes negligible for $R \gg a$, it is assured that the purely quantal properties of the metal are unimportant unless the charge is so close to the metal that surface structure effects are already important.

In the case of a potential due to a point charge $\Delta_1 E$ becomes

$$\Delta_1 E = 0.46 \kappa^{\frac{1}{2}} R^{-1} |W(R)|,$$

and for copper this is

$$\Delta_1 E = 0.89 (\hbar^2/me^2) R^{-1} |W(R)|.$$

It would perhaps be expected that better approximations to $\rho'(\mathbf{r})$, and hence lower values for $\Delta_1 E$, could be made by using smoother functions of z than the step function. A similar calculation using a triangular function in z

$$\rho'(\mathbf{r}) = \frac{2}{\beta^2}(\beta - z)\sigma(s) \quad \text{for} \quad z \leqslant \beta$$

$$\approx 0 \quad \text{for} \quad z \geqslant \beta,$$

results in

$$\beta = (10/3\pi)^{\frac{1}{2}} \kappa^{\frac{1}{2}},$$

$$\Delta_1 E = 2.55 \kappa^{\frac{1}{2}} \int \sigma_0^2(s) dS.$$

Since $(8\pi/3)^{\frac{1}{2}}$ [see Eq. (15)] is 2.89, the energy $\Delta_1 E$ is in fact lowered by this choice of $\rho'(\mathbf{r})$.

For the exponential in z

$$\rho'(\mathbf{r}) = \beta^{-1} e^{z/\beta} \sigma(s),$$

we obtain

$$\beta = (\kappa/2\pi)^{\frac{1}{2}},$$

$$\Delta_1 E = 2.51 \kappa^{\frac{1}{2}} \int \sigma_0^2(s) dS.$$

It should perhaps be mentioned that for some metals (e.g., Bi) the free electronic density n_0 is so small that the plane wave estimate of κ becomes negative. In this case the Fermi-Thomas-Dirac method as used here is inadequate and better wave functions are required to estimate κ.

It has been assumed throughout this discussion that $\rho' \ll \rho_0$. This condition is satisfied for an external charge of the order of e unless it is so close to the metal that surface structure effects have already become important.[1] If a macroscopic point charge, say 10^{10} e, could be brought up to the metal surface, a deviation from classical results would be expected at a distance $R \approx 10^{-8}$ cm. In general the distance R at which deviations are expected is $R \approx Q^{\frac{1}{2}} a$, where Qe is the charge on the macroscopic body assumed to be a sphere of radius r. Then r must necessarily be less than R and the potential V required to charge up the sphere is $Qe/r > Qe/R = (R/a)(e/a) = 27(R/a)$ volts. To observe a macroscopic effect, R/a must be $\gtrsim 10^6$, which requires a potential $V > 10^6$ volts. Thus it seems safe to assume that this effect has no practical importance.

Production of 2537 Radiation and the Role of Metastable Atoms in an Argon-Mercury Discharge

CARL KENTY
Lamp Development Laboratory, General Electric Company, Nela Park, Cleveland, Ohio
(Received July 17, 1950)

The average population of 6^3P_1 mercury atoms is calculated from the measured 2537 output and resonance radiation diffusion theory. Average populations of $6^3P_{2,0}$ are then found by comparing the absorptions of 5461, 4358, and 4047. Populations for a 0.42-amp. discharge in 3.5-mm argon of high purity plus Hg vapor at 42°C in a tube 1½ inches in diameter are for $6^3P_{2,1,0}$ respectively, 7.6×10^{11}, 1.9×10^{11}, and 3.5×10^{11} cm^{-3}; these are ½, 1/20, and ½ of the corresponding Boltzmann populations, for the electron temperature of 11,100°K.

Rates of collisions of the first and second kinds involving electrons (average concentration 2.1×10^{11} cm^{-3}), and the 6^1S_0 and $6^3P_{2,1,0}$ states are calculated using known excitation functions and detailed balancing. The results predict populations for $6^3P_{2,1,0}$ within 10 percent of those observed; they further predict that ⅔ of the observed 2537, ⅔ of the visible triplet, and ¾ of the 3650 group are produced stepwise via 6^3P. Quenching collisions involving argon are assumed negligible.

The results account fairly well for the maximum in the 2537 *versus* Hg pressure curve, for the decrease in the 2537 output efficiency with increasing current, and for the decrease in electron temperature which is associated with both these effects.

INTRODUCTION

THE phosphor in a fluorescent lamp is activated primarily by the resonance radiation of Hg at 2537A. It is a known fact that this radiation does not increase linearly with current as might be expected if excitation and ionization were produced only by single-stage impacts. The present study[1] indicates that two-stage processes involving the resonance and metastable levels of Hg play a far more important role than direct excitation and ionization from the normal state.

In arriving at this conclusion it has been necessary to consider quantitatively the diffusion of resonance radiation, the populations of the $6^3P_{2,1,0}$ states of Hg and the various elementary processes of excitation and de-excitation involving these states.

The discharge studied is a positive column of 420 ma in a 1½-inch tube containing 3½-mm argon plus Hg vapor at $\sim7\mu$ pressure. This is approximately the discharge met with in the standard 40 w fluorescent lamp.

On account of the retarding effect of the argon, few metastable Hg atoms will be able to diffuse to the wall before they are destroyed by collisions with electrons. These collisions will be of a number of types; (1) collisions of the second kind which return the metastable state to the normal state, or in the case of 6^3P_2 to a lower 6^3P state, (2) collisions of the first kind raising a metastable state to a higher excited state, and (3) collisions resulting in ionization of the metastable state. The resonance state 6^3P_1 will suffer similar collisions, but to a smaller extent, owing to its smaller population (as limited by its radiation of 2537). In the attempt to estimate the rates of all these processes it was first necessary to determine the populations of the $6^3P_{2,1,0}$ states with the help of measurements of the relative absorption of the visible triplet lines. These lines, 5461, 4358, and 4047, have 7^3S_1 as a common initial level and end on $6^3P_{2,1,0}$ respectively.

[1] The principal results as stated here were reported at the Columbus Meeting of the Ohio Section of the American Physical Society, March 13, 1948 [Phys. Rev. 74, 114 (1948)] and at the April 1948 M.I.T. Electronics Conference.

ABSORPTION MEASUREMENTS

Absorption measurements were carried out by placing an emission lamp (1) behind an absorption lamp (2) and arranging a slit system requiring that a beam of light originating in (1) pass through (2) before reaching a photo-cell. The slits were 3½-mm wide and 2-cm long, cut in sheets of aluminum foil wrapped around the lamps. Corning monochromatic filters, placed before the cells, were used to isolate the lines. The lamps were placed 1 cm apart. The photo-cell had an opening 1-cm wide and was placed 10 cm in front of lamp (2).

The fractional absorption A of the beam will be given by

$$A=(I_1+I_2-I_{1+2})/I_1, \qquad (1)$$

where I_1 is the intensity of 2537 received when only lamp (1) is on, I_2 the intensity when only lamp (2) is on and I_{1+2} the intensity when both lamps are on. Measurements were made over a current range of 0.1 to 420 ma in the absorption lamp and of 50 to 500 ma in the emission lamp. Results for equal currents in emission and absorption lamps are shown in Fig. 1. Absorptions were read when the wall temperature of the absorption tube was 42°C, as measured by a fine thermocouple on the tube wall. With a current of 420 ma in the emission lamp, its temperature was also 42°C. The lamps were heated solely by their own currents, and in the case of the absorption lamp, the current was dropped from 420 ma only long enough to obtain an absorption reading. In this time the lamp temperature dropped by only a small amount and the absorption, a slowly varying quantity, did not change appreciably.

In Fig. 1, 5461 is seen to be more strongly absorbed than 4047, indicating a higher population for 6^3P_2 than for 6^3P_0. Substantial saturation is seen to set in at

$\sim$100 ma. Since saturation means constancy of population, it is concluded that at $\sim$100 ma the loss of metastables by diffusion to the walls is no longer important as compared with losses by quenching collisions with electrons. Thus beyond $\sim$100 ma the processes of excitation and destruction of metastables, both being proportional to the electron concentration, will increase with current at about the same rate so no great change in population will result.

The absorption of 4358 in Fig. 1 is seen to be the smallest and to show distinctly less saturation than the others. This is due to the fact the loss of 6^3P_1 by radiation is still high compared with loss by collisions of the second kind with electrons.

Regarding the saturation of the metastable populations at low current, it is instructive to consider the time taken for a metastable to reach the walls. Thus with a mean free path of $\sim$0.006 cm, and a speed of $\sim 1.8\times10^4$ cm sec.$^{-1}$, a metastable will require a time of the order of

$$(1.8/0.006)^2\times0.006/(1.8\times10^4)\cong0.03 \text{ sec.},$$

to reach the wall of a tube 1.8 cm in radius. Calculations show that this is ample time for dissipative collisions with electrons to occur, even with small currents. This means that the distribution of metastables over the cross section of the lamp will be much flatter than that of the resonance atoms as Fabrikant[2] has shown.

The gas in the lamps contained <2 parts per million of impurities and quenching of the 6^3P Hg states by them was negligible. Complete evidence for this, and further evidence concerning the saturation of the metastable populations, resulting from an extended study of these discharges at low currents will be given in a later paper. Quenching of metastables by argon is negligible in comparison to quenching by electrons (cf. the work of Panevkin, to be discussed further below).

While the absorption data were obtained with a d.c. discharge, and the 2537 and visible triplet data were obtained with the usual a.c. discharge, this is probably not of great importance here since within the present accuracy the a.c. and d.c. spectral characteristics are similar.[3]

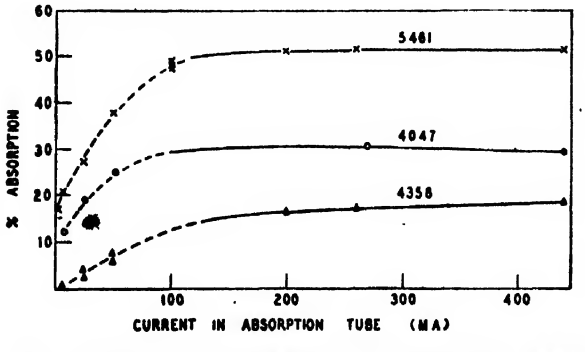


FIG. 1. Absorption of the visible triplet lines of Hg in passing through a d.c. discharge in a mixture of 3.5-mm $A+8\mu$ Hg vapor in a 3.6-cm diameter tube. Emission lamp is a similar tube. Above 100 ma similar currents are used in both tubes for each experimental point. At the higher currents the Hg pressure in the emission tube is also about 8μ.

[2] V. A. Fabrikant, Comptes Rendus U.S.S.R. 19, 385 (1938).

CALCULATION OF THE RELATIVE POPULATIONS OF THE $6^3P_{2,1,0}$ STATES FROM THE ABSORPTIONS

If in the above absorption study the line radiation passed through the absorbing gas unchanged in nature, the number of absorbing atoms would be proportional to A, for small absorptions, and in general to $\log(1/1-A)$. The values of $\log(1/1-A)$ for 0.42 amp have been evaluated, and their ratios are given in Table I. But the core of the line is relatively the most absorbed and so the line is broadened as it passes through the discharge. Also the breadths of the lines from the emission tube will in general be different. There is one case however solved by Ladenburg and Reiche,[4] where k_0, the absorption coefficient of the core of the line can be deduced from the measured A; this is the case where conditions in the emission and absorption tubes are the same. It was for this reason that care was taken to have currents and temperatures the same at 420 ma in Fig. 1. The value of k_0 for the different lines has been evaluated, using the relation of Ladenburg and Reiche, from the A values of Fig. 1 and are given in Table I. The procedure has followed that of Fabrikant, Butaeva, and Zirg.[5] It has been stated by these authors that the atomic absorption coefficients for the lines of the visible triplet are all equal; this can be proved by application of the sum rules and the Einstein theory of radiation. Consequently the average concentrations of the absorbing atoms will be proportional to the k_0's in Table I, where the ratios are given in the last column.

It will be seen that the k_0's give somewhat smaller relative populations to 6^3P_1 and 6^3P_0 than do the values of $\log(1/1-A)$. This is to be expected since 4358, for example, is relatively sharper coming from the emission lamp than 5461 and, therefore, is more highly absorbed, per absorbing atom.

CALCULATION OF THE POPULATION OF 6^3P_1 ATOMS FROM THE 2537 OUTPUT

Thayer and Barnes[6] determined that in the present lamp operating on a.c. at 420 ma in 3.5-mm A plus Hg at 7.6μ (43°C), about 60 percent of the 40w input, or 24 watts, is converted into radiation able to excite the phosphor. According to further unreported work by the same authors in this laboratory, probably 2 watts of this is 1849 radiation; 22 watts is therefore taken as the 2537 output. Division of this by the length of the dis-

[3] W. E. Forsythe and E. Q. Adams, *Fluorescent and Other Gaseous Discharge Lamps* (Murray Hill Book Company, New York, 1948), p. 152.

[4] See A. C. G. Mitchell and M. W. Zemansky, *Resonance Radiation and Excited Atoms* (Cambridge University Press, London, 1934), p. 118 ff. and Table p. 323.

[5] Fabrikant, Butaeva, and Zirg, Physik. Zeits. Sowjetunion 11, 576 (1937); 13, 23 (1938).

[6] R. N. Thayer, Trans. Electrochem. Soc. 87, 413 (1945).

charge, about 110 cm, between electrodes, and by the energy equivalent of the 2537 quantum (7.78×10^{-19} joule) gives the rate of escape R of 2537 quanta to be 2.5×10^{17} per cm length per second. Assuming that net quenching losses of the 6^3P_1 state are small, and for simplicity that the radial distribution of 6^3P_1 atom concentration n_1 will be roughly the parabolic one

$$n_1=(n_1)_0[1-(r^2/a^2)], \tag{2}$$

where $(n_1)_0$ is the concentration at the center and a is the tube radius, the concentration gradient at the wall will be $-2(n_1)_0/a$ and this multiplied by the diffusion coefficient of the radiation D and the circumference of the tube $2\pi a$ should give the rate of escape. Thus

$$R=4\pi(n_1)_0D. \tag{3}$$

The diffusion coefficient used is the approximate one arrived at by the author[7] and put into more convenient form by Zemansky.[8] Namely,

$$D=1/(4\bar{k}^2\tau). \tag{4}$$

Here τ is the mean life of the 6^3P_1 state (1.08×10^{-7} sec.) and $\bar{k}^2$ is the effective average square absorption coefficient. Tables for $\bar{k}$ the r.m.s. or "equivalent" absorption coefficient are given by Zemansky[9] in the form of $\bar{k}l$ as a function of k_0 where l is the thickness of the absorption cell and k_0 is the absorption coefficient of the core of the line. For the present case $k_0=n_ak_0'$ where k_0' is the atomic absorption coefficient for the core of the line and n_a is the concentration of Hg atoms. Substitution[10] of $k_0'=1.21\times10^{13}$, $n_a=2.13\times10^{14}$ cm^{-3} and $l\cong4/3a$ or 2.4 cm, gives $k_0=27$, $\bar{k}=6.7$ and $D=5.2\times10^4$.

Substitution for D and R in Eq. (3) then yields for $(n_1)_0$ the value 3.8×10^{11} cm^{-3}, an average value of n_1 over the cross section of 1.9×10^{11} cm^{-3}, and a total number N in a 1-cm length of tube (cross section ~10 cm^2) of 1.9×10^{12}. To account for the observed output, $R=2.5\times10^{17}$ per sec. per cm, the effective life T of the 2537 quantum within the tube would need to be $N/R=0.76\times10^{-5}$ sec.

The reciprocal of $\bar{k}$ is the effective mean free path or 0.15 cm. According to the principle of the random walk, the number of absorptions and re-emissions suffered by a quantum generated at the center before escape will be of the order of $(a/\lambda)^2$, i.e., $(1.8/0.15)^2\cong140$. The average for all quanta will be less than this; half the above number or 70 would be closely in accord with value of effective life of 0.76×10^{-5} sec. arrived at above.[11]

[7] C. Kenty, Phys. Rev. 42, 823 (1932).
[8] M. W. Zemansky, Phys. Rev. 42, 843 (1932).
[9] See also reference 4, Table XII, p. 331.
[10] k_0' has been calculated from Garrett's value of τ of 1.08×10^{-7} sec., following Zemansky, Phys. Rev. 36, 229 (1930). n_a has been calculated for 7.6μ Hg pressure on an average gas temperature in the lamp of 73°C (346°K) as found in work in this laboratory to be reported elsewhere.
[11] V. A. Fabrikant, Comptes Rendus U.S.S.R. 19, 389 (1938) gives the approximate expression $3k_0a\tau$ for this effective life. With $k_0=26$, $a=1.8$ cm and $\tau=1.08$ sec. This expression gives $T=1.5\times10^{-5}$ sec. which is twice the value obtained here.

TABLE I. Absorption of the visible triplet. A, absorption taken from Fig. 1; k_0, absorption coefficient for core of line as obtained from Ladenburg and Reiche's relation between k_0l and A (reference 3 l has been taken as 3.6 cm).

Line	A	Relative $\log(1/1-A)$	k_0	Relative k_0
5461	0.51	100	0.32	100
4358	0.18	28	0.083	25
4047	0.30	49	0.15	46

The problem may be attacked alternatively using the more recent and exact radiation diffusion theory of Holstein.[12] According to this theory, the distribution of 6^3P_1 atoms across a tube quickly assumes a characteristic form after the excitation, in whatever manner, is cut off, and the 2537 emitted thereafter decays with a characteristic time T given by[13]

$$T=\tfrac{5}{8}k_0a\tau(\pi\log k_0a)^{\frac{1}{2}}, \tag{5}$$

where a is the tube radius. Substitution of the values of the constants already used yields $T=0.72\times10^{-5}$ sec. This value of the decay time, or average life, is seen to be only slightly smaller than the average life obtained by the first method (0.76×10^{-5} sec.); which attests to the essential equivalence of the two methods for a case like the present. The first method suffers from the arbitrariness of the choices of the parabolic distribution and of the value of $l=(4/3)a$ as the thickness of the equivalent absorption cell. In the second method, the formula used applies to the characteristic distribution of 6^3P_1 atoms and not necessarily to the distribution existing in the actual discharge. In any case, since the practical difference in the two methods appears to be so small, the results will be left in their original form as based on $T=0.76\times10^{-5}$ sec.

THE POPULATIONS OF THE $6^3P_{2,1,0}$ STATES

If there were thermal equilibrium between the 6^3P states and the electrons at their temperature T_e, their populations would be given by

$$n=n_a(g/g_a)e^{-(eV_e)/(kT_e)}, \tag{6}$$

where $n_a=2.13\times10^{14}$ is the concentration of normal atoms, g_a is the statistical weight of the normal atom, g is the weight of the 6^3P atom in question, and V_e is its excitation potential. For T_e the value of 11,100°K is used, based on probe measurements made in this laboratory by M. A. Easley.[14] The populations as thus calculated are given in Table II. The last column of the table gives the ratios of the populations calculated from absorption to the Boltzmann populations.

[12] T. Holstein, Phys. Rev. 72, 1212 (1947); see also L. M. Bieberman, J. Exp. Theor. Phys. 17, 416 (1947).
[13] Alpert, McCoubrey, and Holstein, Phys. Rev. 76, 1257 (1949). The formula as quoted in their article inadvertently omits the factor τ.
[14] The author is indebted to Miss Easley for permission to use her unpublished values for both T_e and n_e.

TABLE II. Comparison of populations with Boltzmann populations. V_e, excitation potential; g, statistical weight; k_0, absorption coefficient of core of line (from Table I); n, calculated from the ratio of the k_0's and $n=1.9\times10^{11}$ cm^{-3} for 6^3P_1, obtained from output and diffusion of 2537; n_B, Boltzmann population for $T_e=11{,}100°$K.

State	V_e	g	Ratio k_0	$n\times10^{11}$	$n_B\times10^{11}$	n/n_B
6^3P_2	5.43	5	100	7.6	36.4	0.21
6^3P_1	4.86	3	25	1.9	40.2	0.047
6^3P_0	4.66	1	46	3.5	16.3	0.21

The use of a temperature here for the electrons is taken to be justified even though the distribution may differ from Maxwellian in being deficient in the very high speed members; for the experiments of Miss Easley indicate that such deficiency occurs only for electrons of considerably higher energy than is here involved.

Fabrikant, Butaeva, and Zirg[15] calculated absolute values for the populations of $6^3P_{2,1,0}$ directly from their absorption measurements. They deduced k_0 in each case from the measured absorption by use of the Ladenburg-Reiche formula, as was done above, and used known relations[16] connecting the k_0's with τ, the mean life of the 7^3S_1 state. For τ they used Mitchell and Murphy's value[17] of $0.8\pm0.1\times10^{-8}$ sec. They assumed Doppler broadening and that therefore (with gas atoms at substantially room temperature) the five hyperfine structure components did not overlap, so that the number of absorbing atoms was $\frac{1}{5}$th the number in the state concerned. If the same procedure is followed here, populations are obtained which are about a third those found above. Panevkin[18] following the same procedure as Fabrikant, Butaeva, and Zirg, and studying a discharge rather similar to the present one, has obtained $6^3P_{2,1,0}$ populations one-third to one-fourth as large as those found here. It is believed that the optical densities involved are too great for such application (in the absolute) of the Ladenburg-Reiche formula, the values of k_0 derived with it being too small. Such an error becomes very pronounced[19] if an attempt is made to apply this method to the case of the 1-At. Hg arc where the densities of the 6^3P atoms are about 100-fold higher. Since the populations in Fabrikant, Butaeva, and Zirg's discharge were of the same order as those found here, it is believed their estimates are too low, perhaps by a factor of 2 or 3. In accord with this is the fact that they need populations several-fold higher than they calculate, in order to explain their observed 2537 output.[20]

QUASI EQUILIBRIUM FOR THE METASTABLE STATES

It was at first thought that the flatness of the absorption curves for these states in Fig. 1 indicated a true state of thermal equilibrium with the electrons. However, Table I does not indicate this. Further, following Panevkin,[21] true thermal equilibrium in a case like this would require, not constant populations with regard to current increase, but decreasing ones due to the fact that T_e decreases with current. It is concluded therefore that the absorption curves represent a state of quasi equilibrium with the electrons in which detailed balancing is not yet complete. Thus the populations of the metastable states are determined solely by electron collisions and the balance (not yet a Boltzmann one) changes only slowly with the current. All 6^3P states are closely tied together through collisions of the first and second kinds as Yavorsky[22] has shown; as will appear below, 6^3P_2 and 6^3P_0 are held below Boltzmann populations because of the important "leakage" of 2537.

The flatness of the absorption curves, considering the falling T_e, really means a gradual approach toward true thermal equilibrium conditions in which $^3P_2$ for example would be quenched to normal by electrons as fast as

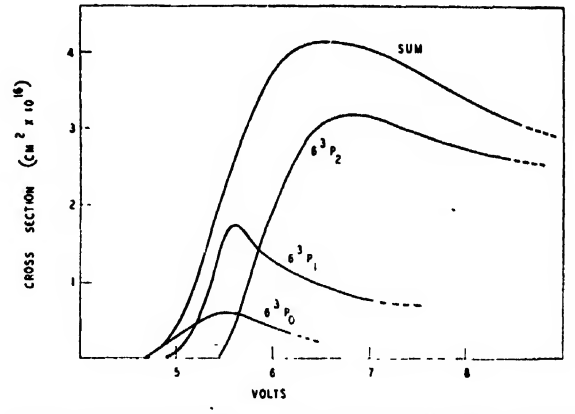


FIG. 2. Excitation function for the 6^3P states of Hg. That for 6^3P_1 is Nottingham's photoelectrically determined curve. The relative heights of the maxima of the three curves and also as far as possible their relative shapes are in accord with the quantum mechanical calculations of Penney and of Yavorsky. The scale of the three curves and some details of their shapes are chosen so that their sum reproduces the curve of Arnot and Baines, obtained from purely electrical measurements.

[15] Fabrikant, Butaeva, and Zirg, Physik. Zeits. Sowjetunion **11**, 576 (1937) and other papers in this series.

[16] See reference 4, p. 101 ff., where $1/\tau$ of Eq. (35) is to be replaced by the Einstein transition probability A, for the line in question and the measured τ of the 7^3S_1 state is the reciprocal of the sum of the A's for the green blue and violet lines. The relation between the A's being given by the sum rules which state that $A\propto g\nu^3$, g being the statistical weight of the lower state concerned, and ν being the frequency of the line.

[17] A. C. G. Mitchell and J. M. Murphy, Phys. Rev. **46**, 53 (1934).

[18] K. Panevkin, Papers of Acad. of Sci. U.S.S.R. **59**, 683 (1948).

[19] Using Barnes and Adams' [Phys. Rev. **53**, 545 (1937)] measurements of the absorption of 5461, 4358, and 4047 for a $\frac{1}{4}$-atmosphere pressure Hg arc, the same method has yielded populations for the 6^3P states which are at least 25-fold smaller than those calculated from Boltzmann's theorem and the temperature as determined by x-rays [C. Kenty and W. J. Karash, Phys. Rev. **78**, 625 (1950)].

[20] While the authors seem to obtain an agreement, they base it on a speed of escape of 2537 through the Hg vapor which is much faster than that allowed by the present treatment of radiation diffusion or by Fabrikant's own (presumably) later treatment (reference 11).

[21] K. Panevkin, J. Phys. U.S.S.R. **2**, 39 (1940). See also V. Fabrikant and K. Panevkin, Comptes Rendus U.S.S.R. **20**, 441 (1938).

[22] B. Yavorsky, J. Phys. U.S.S.R. **10**, 476 (1946).

excited from normal and all other processes would be balanced by their opposites.

COMPUTATION OF THE RATES OF EXCITATION OF THE $6^3P_{2,1,0}$ STATES

While existing knowledge of the excitation functions for the 6^3P states of Hg leaves much to be desired, the following calculations, based on what are judged to be the best of the evidence, yield results which are in remarkably good agreement with the observed 2537 output. This agreement is believed to lend substantial support to the essential validity of the functions chosen. For 6^3P_1 Nottingham's photoelectric relative excitation function[23] has been used. For 6^3P_0 and 6^3P_2, approximate relative shapes for the functions have been taken from Penney's[24] and Yavorsky's quantum mechanical calculations.[25] The functions chosen are shown in Fig. 2. The relative heights of the three maxima are approximately also in shape with that found by Arnot and Baines[26] from purely electrical measurements of inelastic cross sections.

This procedure gives a maximum for 6^3P_1 substantially in agreement with Bricout's optically determined value of 6 cm² cm⁻³ or 1.7×10^{-16} cm² as quoted by Brode.[27,28] (The latter number is obtained from 6 cm² cm⁻³ by dividing by 3.56×10^{16}, the number of molecules in 1 cm³ of gas at 1 mm pressure and 0°C.) All cross section maxima used or deduced in the present study are assembled in Table III.

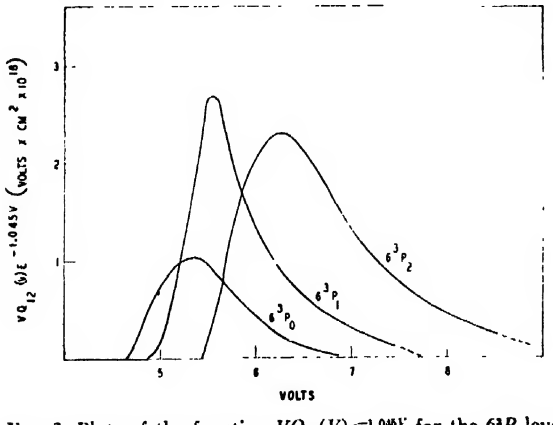


FIG. 3. Plots of the function $VQ_{12}(V)\epsilon^{-1.045V}$ for the 6^3P levels of Hg. $Q_{12}(V)$ is the cross section in cm² for the excitation of the level by electrons of energy V volts. The areas under the curves are proportional to the rates of excitation of the levels.

[23] W. B. Nottingham, Phys. Rev. **55**, 203 (1939).
[24] W. G. Penney, Phys. Rev. **39**, 467 (1932).
[25] B. Yavorsky, Bull. Acad. Sci. U.S.S.R. **9**, 233 (1945). While Penney has given only relative values, Yavorsky has calculated absolute ones which are stated to be in accord with the results of Arnot and Baines. Both authors agree fairly well as to the relative heights of the maxima, but Yavorsky's maxima occur at considerably higher voltages, and the shapes are otherwise somewhat different from Penney's.
[26] F. Arnot and O. Baines, Proc. Roy. Soc. **A151**, 256 (1935).
[27] R. B. Brode, Rev. Mod. Phys. **5**, 257 (1933).
[28] The cross section per atom is obtained from the total cross section per cm³ at 1 mm pressure by dividing by 3.56×10^{16}, i.e., the concentration of molecules at 1 mm and 0°C.

TABLE III. Cross section maxima per atom used or implied in the present study; $Q_{12}(V)$ for electron collisions of the first kind, $Q_{21}(V)$ for electronic collisions of the second kind.

Transition	Threshold (volts)	$Q_{12}(V)$ max. cm² $\times10^{16}$	$Q_{21}(V)$ max. cm² $\times10^{16}$	Reference
$6^1S_0\rightleftarrows6^3P_2$	5.43	3.2	4.1	Nottingham,* Penney,** Yavorsky***
$6^1S_0\rightleftarrows6^3P_1$	4.86	1.74	4.6	Arnot and Baines†
$6^1S_0\rightleftarrows6^3P_0$	4.66	0.58	4.5	Bricout†† (see text)
$6^1S_0\rightarrow7^3S_1$	7.69	0.45		Yavorsky†††
$6^3P_2\rightarrow7^3S_1$	2.26	~4.0		Fabrikant, Butaeva and Zirg‡ (see text)
$6^3P_1\rightarrow7^3S_1$	2.83	~4.00		
$6^3P_0\rightarrow7^3S_1$	3.04	3.9		Yavorsky***
$6^3P_1\rightleftarrows6^3P_2$	0.57	19.3	44.0	Yavorsky‡‡
$6^3P_0\rightleftarrows6^3P_1$	0.20	(~15.0)	(~58.0)	(See text)

* See reference 23.
** See reference 24.
*** See reference 25.
† See reference 26.
†† See reference 27.
††† See reference 30.
‡ See reference 5.
‡‡ See reference 22.

Assuming a Maxwellian distribution for the electrons, the rate of excitation Y per cm of length will be given by

$$Y=2.66\times10^{14}\frac{a^2n_en_a}{T_e}\int_{V_e}^{\infty}VQ_{12}(V)\epsilon^{-(eV/kT_e)}dV, \quad (7)$$

where $Q_{12}(V)$ is the excitation cross section in cm² as a function of the electron energy V in volts and n_e is the electron concentration averaged across the tube (2.1×10^{11} cm² according to Easley's[14] probe measurements). With $a=1.8$ cm and $T_e=11{,}100$°K, Eq. (7) becomes

$$Y=1.64\times10^{20}n_a\int_{V_e}^{\infty}VQ_{12}(V)\epsilon^{-1.045V}dV. \quad (8)$$

Plots of the quantity $VQ_{12}(V)\epsilon^{-1.045V}$ are given in Fig. 3. The areas under these curves were found graphically and taking $n_a=2.13\times10^{14}$ cm⁻³, the rates of excitation were found to be 1.35×10^{17}, 0.90×10^{17} and 0.40×10^{17} for the $6^3P_{2,1,0}$ states respectively.

An attempt has been made in Fig. 4 to portray the various transitions. Here the thickness of the 6^3P levels have been drawn pictorially in the ratios of the populations, i.e., calculated n values in Table II. Transitions by collisions of the first kind (excitation by electrons) are represented by straight arrows pointed upward, and collisions of the second kind (de-excitation by electrons) by straight arrows pointed downward. Radiations are represented by wavy arrows. A strength of 3×10^{16} transitions per sec. per cm length has been given to full arrows (either straight or wavy). Dotted arrows have the strengths indicated in decimals of the same unit. The relative total rate of any transition may then be visualized by adding the number of full arrows involved and the decimal part of an arrow. Thus the rates of excitation from normal of the $6^3P_{2,1,0}$ states calculated above are represented by 4.5, 3.0, and 1.33 arrows respectively, and the absolute rates may be found by multiplying these numbers by 3×10^{16}. Wavy arrows represent *measured* outputs in all cases.

QUENCHING TO THE NORMAL STATE

In complete thermal equilibrium with the electrons at their temperature T_e, quenching of any level to the normal state by electrons would, by the principle of detailed balancing, take place at the same rate as excitations to that level from the normal state. Since the rate of quenching will be proportional to the population of the upper state in question, and since the populations of the metastable states have each been estimated as 0.21 times the Boltzmann populations, their rates of quenching to normal by electrons will be 0.21 times their rates of excitation from normal, which rates have been calculated above. Likewise the rate of quenching of 6^3P_1 to normal will be 0.047 times the rate of excitation from normal. These quenching rates are shown by downward arrows in Fig. 4; 0.95 arrow for 6^3P_2, 0.14 arrow for 6^3P_1, and 0.28 arrow for 6^3P_0. The sum of these, 1.37, is 16 percent of the total excitations from normal.

Calculation of the actual quenching cross section curves[29] brings out the interesting circumstance that the maxima of these curves are all about equal (see Table III), while the Q_{12} maxima are approximately in the ratio of the g's.

The escaping 2537 has been represented in Fig. 4 as 8.3 wavy arrows to represent the measured 2537 output. The heavy wavy arrows represent the absorptions and re-emissions of 2537 before they escape; they have a strength each of the order of $\sim 70 \times 8.3$ or ~ 580 of the other full arrows. Their effect is merely to increase about 70-fold the effective lifetime of the 6^3P_1 state, a circumstance which has already been taken account of in calculating the population of this state.

TRANSFERS AMONG THE $6^3P_{2,1,0}$ STATES

The electronic excitation function for the $6^3P_1 \rightarrow 6^3P_2$ transition has been calculated from quantum mechanics by Yavorsky.[22] This is shown in Fig. 5. Again using Eq. (8) and n (in place of n_a) as 1.9×10^{11} cm^{-3} for the 6^3P_1 state, the rate of such excitations has been calculated to be 3.27×10^{16} cm^{-1} sec.$^{-1}$. This is represented by 1.09 upward arrows in Fig. 4. Figure 5 also shows the cross section $Q_{21}(V)$ for collisions of the second kind[29] quenching 6^3P_2 to 6^3P_1.

To calculate the rate of quenching transfers 6^3P_2

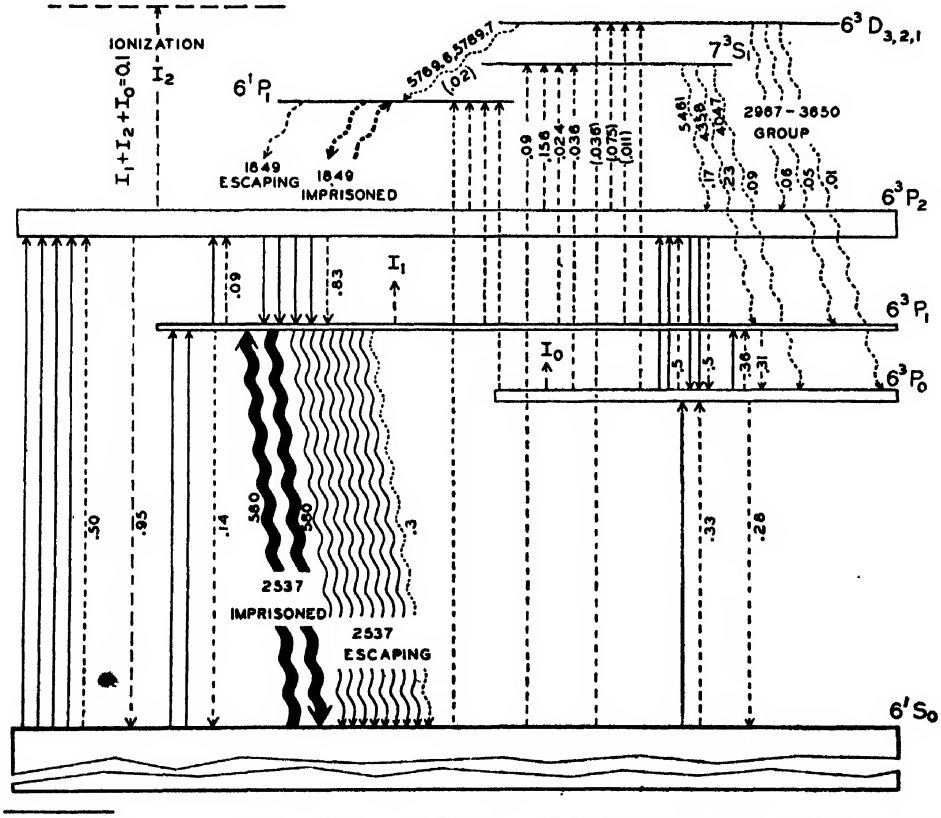


FIG. 4. Principal processes occurring in the fluorescent lamp. Straight arrows are calculated collisions of the first and second kinds. Wavy arrows are measured radiation outputs (except the heavy arrows). One full arrow, either straight or wavy represents 3×10^{16} transitions per sec. per cm of length. Decimals are fractions of full arrows. (*Correction added in proof:* There should be three full upward arrows from $6'S_0$ to 6^3P_1 instead of two as shown.)

[29] Quenching curves, $Q_{21}(V)$, have been calculated from the corresponding $Q_{12}(V)$ curves using the Klein-Rosseland relationship

$$Q_{21}(V) = \frac{g_1}{g_2} \frac{V}{V - V_e} Q_{12}(V),$$

where it is to be understood that V in $Q_{21}(V)$ on the left hand side is equal to $V - V_e$ on the right-hand side, and g_2 and g_1 are the statistical weights of the upper and lower states respectively. However, as seen above, it is not necessary to use these curves explicitly, a fact which seems not to be generally recognized.

$\rightarrow 6^3P_1$, it is first noted that if 6^3P_1 had a full Boltzmann population, the upward transition rate ($6^3P_1 \rightarrow 6^3P_2$) would be $(1/0.047) \cong 21$ times as great as it is, other things being equal, since 0.047 is the fractional Boltzmann population of 6^3P_1 (Table II). This rate of transitions (69×10^{16} sec.$^{-1}$ cm^{-1}) would also be the rate of downward ($6^3P_2 \rightarrow 6^3P_1$) transitions if 6^3P_2 had its full Boltzmann population. But since its population is actually 0.21 times this, the downward rate will be $0.21 \times 69 \times 10^{16} = 14.5 \times 10^{16}$ sec.$^{-1}$ cm^{-1}. This is shown as 4.83 downward arrows in Fig. 4. To calculate the rate of transitions in one direction from the known rate in the other direction, it is therefore only necessary to multiply by the ratio of the fractional Boltzmann populations.

Unfortunately no information exists regarding cross sections for transfers between 6^3P_0 and $6^3P_{2,1}$ states. But certain inferences may be drawn on the basis of the scheme of Fig. 4. Thus both 6^3P_2 and 6^3P_0 have, in this scheme, one quarter the Boltzmann population concerned. Transfers between the two states up and down must therefore take place at the same rate, just as they would in complete thermal equilibrium (when they would be ~five times as fast). Whatever the rate of the transfers, they will therefore not significantly change the scheme of Fig. 4 as drawn. The allotment of $2\frac{1}{2}$ arrows up and $2\frac{1}{2}$ down is arbitrary, but is compatible with a cross section for $6^3P_0 \rightarrow 6^3P_2$ of the same order as that for the $6^3P_1 \rightarrow 6^3P_2$ transfer and is in accord with evidence drawn from low current studies to be reported elsewhere. Here as elsewhere it is assumed that the number of transfers brought about by impacts of atoms or ions is negligible.[29a]

It is possible to draw more definite information regarding the transitions $6^3P_0 \rightarrow 6^3P_1$ and their reverse. Thus, in order to maintain 6^3P_0 at the constant population shown, there must be in total as many arrows leaving as ending on it. Now since a net gain of 1.05 arrows, as above defined, has already been calculated for it involving 6^1S_0, and since transfers to and from 6^3P_2 are equal and no other important losses or gains are involved, as will be seen, transfers involving $6^3P_0 \rightleftarrows 6^3P_1$ must be chosen so as to give a net gain of 1.05 arrows to 6^3P_1 from 6^3P_0. Let x be the upward rate and y the downward rate. Then $x-y=1.05$ and $x/y=0.21/0.047$ (the ratio of the fractional Boltzmann populations of Table II). Solution of these equations gives $x=1.36$, $y=0.31$. This selection of x and y would result if the cross section for $6^3P_0 \rightarrow 6^3P_1$ were 75 percent of that for 6^3P_1 and 6^3P_2. It thus appears that the cross sections for excitation from one 6^3P state to another are not greatly different from one another.

TRANSFERS INVOLVING 7^3S_1

Fabrikant, Butaeva, and Zirg[5] have concluded from their experiments that the cross sections for excitation

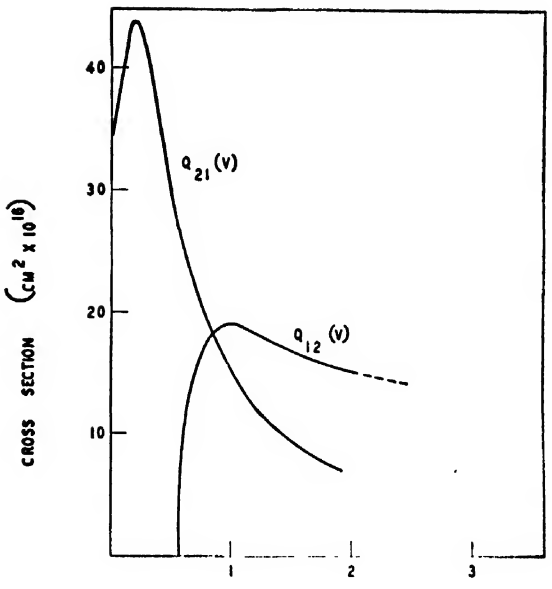


FIG. 5. $Q_{12}(V)$: Cross section for electron collisions exciting 6^3P_2 from 6^3P_1 (Yavorsky, reference 22). $Q_{21}(V)$: Cross section for collisions of the second kind with electrons returning 6^3P_2 to 6^3P_1. Yavorsky expresses his cross sections in the atomic units $10^{16}\pi a_0^2$ (=0.87 cm^2) per cm^3 at one mm pressure of the atoms concerned, where a_0 is the radius of the first Bohr orbit of the H atom (0.528×10^{-8} cm). The conversion factor is therefore $0.87/3.56 \times 10^{16} = 2.45 \times 10^{-17}$, so that his maximum of 78 in his units equals 19.4×10^{-16} cm^2.

of 7^3S_1 from $6^3P_{2,1,0}$ are an order of magnitude higher than that for excitation of 7^3S_1 from 6^1S_0. Their estimate was based on the observed faster-than-linear increase of the visible triplet lines with current in a pure Hg discharge at constant pressure. They had estimates of the concentrations of $6^3P_{2,1,0}$ from absorption measurements and knew n_e and T_e (the latter being assumed constant as the current varied). Having estimated the fraction of the visible triplet output due to cumulative excitation at any current, they then calculated excitation cross sections assuming that the shapes of the functions were similar to that for 6^1P_1 excitation from the normal state (broad maxima since no change in multiplicity is involved), and that the functions for all the 6^3P states were the same in magnitude at any given voltage above threshold. The resulting cross sections were 12.5 cm^2 per cm^3 (3.5×10^{-16} cm^2) at the maxima.

Yavorsky[30] has more recently calculated quantum mechanically $Q_{12(V)}$ for $6^3P_0 \rightarrow 7^3S_1$ and $6'S_0 \rightarrow 7^3S_1$. His curves are shown in Fig. 6. The maximum for $6^3P_0 \rightarrow 7^3S_1$ is seen to be 3.9×10^{-16} cm^2 (in rather good accord with Fabrikant, Butaeva, and Zirg's results) and that for $6^1S_0 \rightarrow 7^3S_1$ to be 0.45×10^{-16} cm^2. The latter is also in quite good accord with the value 0.64×10^{-16} cm^2 (1.2 cm^2 per cm^3) obtained earlier by Hanle and Schaffernicht.[31] The maximum in the case of $6^3P_0 \rightarrow 7^3S_1$ is seen to be relatively very broad.

[29a] Some of the evidence for this is summarized by C. Kenty, Phys. Rev. 42, 823 (1932). See especially pp. 838, 839.

[30] B. Yavorsky, Comptes Rendus U.S.S.R. 48, 175 (1945).

[31] W. Hanle and W. Schaffernicht, Ann. d. Physik 6, 905 (1930).

The upward arrows in Fig. 4 have been calculated using the curves of Fig. 6 and the concentrations of $6^3P_{2,1,0}$ given in Table II, and assuming that the cross sections for all three 6^3P states are the same. These add up to a total of 0.31 arrow or 0.93×10^{16} transitions sec.$^{-1}$ cm^{-1}. This indicates that about two thirds of the excitation of 7^3S_1 comes via the $6^3P_{2,1,0}$ states.

THE OUTPUT OF 5461, 4358 AND 4047

The 7^3S_1 state will radiate these lines in the ratio 100:117:49, i.e., in the ratio $g\nu^3$. However, these ratios will be modified somewhat by absorption. Thus 5461, being more strongly absorbed will be weakened relatively to the others and particularly with respect to 4358 which is least absorbed. It is to be noted however that the number of quanta emitted is not changed by absorption since the population of 7^3S_1 is too small for detectable quenching. The calculated excitation rate of 7^3S_1 may thus be compared with the measured rate of emission of the visible triplet.

According to Forsythe and Adams,[32] the outputs of 5461, 4358, and 4047 from the lamp under consideration (40 watt) amount to 0.6 percent, 1.05 percent, and 0.45 percent respectively, of the positive column input (about 33 watts). Converting these to quanta per sec. per cm length one obtains 0.50×10^{16}, 0.70×10^{16}, and 0.28×10^{16}, respectively (corresponding wavy arrows shown in Fig. 4). The sum of these is 1.48×10^{16} sec.$^{-1}$ cm^{-1} and it is seen that the calculated excitation rate of 7^3S_1 (0.93×10^{16} sec.$^{-1}$ cm^{-1}) falls short of this by 37 percent.

The degree of accord is considered good in view of the assumptions made. The disagreement might be accounted for if (a) Yavorsky's excitation function for $6^3P_0\rightarrow7^3S_1$ does not rise sharply enough from onset, or (b) the assumption of the equality of the excitation cross sections for all three transitions $6^3P_{2,1,0}\rightarrow7^3S_1$ is incorrect, or if (a) and (b) are both partly true. (a) seems to be a possibility since in the case of the excitation functions for the 6^3P states from normal, Yavorsky's theoretical curves[25] do not rise nearly as sharply from the onset as the experiments of Nottingham or Arnot and Baines would indicate. In regard to (b) it might be noted that if the cross section for $6^3P_2\rightarrow7^3S_1$ were increased by a factor of two, the discrepancy would be removed.

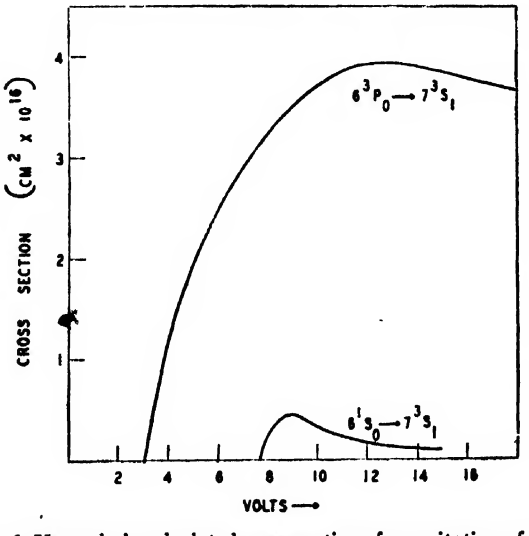


FIG. 6. Yavorsky's calculated cross sections for excitation of 7^3S_1 from 6^1S_0. Reproduced from reference 30, p. 177.

As to the experimental cross sections of Fabrikant, Butaeva, and Zirg,[5] which agree quite well with Yavorsky's[30] results, at least as to the heights of the maxima, it may be that they have, in their calculations, used a function which rises more steeply from onset than does Yavorsky's in order to account for the output. While their estimate seems to be quite good, it is believed that it involves two rather large errors which substantially neutralize each other: (1) Their estimates of 6^3P populations, based on absorption, are, according to the views mentioned above, two- or threefold too low; while (2) the fraction of the visible triplet taken by them to be due to cumulative acts is probably too small by a factor of two or more because of their assumption of the constancy of T_e as the current in their discharge in 1μ pure Hg vapor is raised. Thus their measured 2537 output per electron is markedly decreasing with rising current, showing a falling T_e, and since the direct excitation of 7^3S_1 per electron should decrease faster than that of $6^3P_{2,1,0}$ (i.e., than the 2537 output), it follows that the remainder of the 7^3S_1 excitation, namely that due to cumulative acts, must be greater than they have assumed. Their curves indicate that an error of a factor of two could thus easily have been made. In any case it seems likely that at least 67 percent of the excitation of 7^3S_1 in the fluorescent lamp under consideration comes stepwise via the 6^3P states.

TRANSFERS INVOLVING $6^3D_{3,2,1}$

Cross sections for the excitation of the $6^3D_{3,2,1}$ states (treated as one) have been roughly estimated from the data given by Forsythe and Adams.[32] These authors give as the output of the 40 watt fluorescent lamp the following: for the three lines ending on 6^3P_2 (called 3654), 0.32 percent; for the two lines ending on 6^3P_1 (called 3121), 0.37 percent; and for the yellow lines together, 0.12 percent of the positive column wattage (33 watts). A little over half of the energy of the yellow lines may be taken as coming from 6^3D, i.e., about 0.07 percent. Also the energy of the single line ending on 6^3P_0 (2967) may be taken[33] as 25 percent of that of the 3654 group or 0.08 percent. These percentages have been converted into quanta per sec. per cm length of lamp and are represented by the arrow strengths shown in Fig. 4. The total of the transitions amounts to 4.3×10^{15} sec.$^{-1}$ cm^{-1} or 0.14 arrow. It has been found

[32] W. E. Forsythe and E. Q. Adams, reference 3, Table 11, p. 86.

[33] Private information of Dr. B. T. Barnes.

that this total rate of excitation of 6^3D can be accounted for on the basis of excitation cross section curves, the same as were used for calculating 7^3S_1 transitions but increased 10 percent in magnitude. According to these curves the excitation of 6^3D is distributed as indicated by the arrows in Fig. 4, 75 percent of it coming from the 6^3P states.

TRANSFERS FROM $6^3P_{2,1,0}$ TO 6^1P_1

No data exist for estimating the transfers from the 6^3P states to 6^1P_1. If the cross sections were the same as for the transfers 6^3P to 7^3S_1, a rough calculation shows that transfers would take place about twice as fast to 6^1P_1 as to 7^3S_1. This would give such transfers an importance too small to upset the balance of Fig. 4 appreciably. That such transfers are relatively few is also indicated by the work of Butaeva and Fabrikant[34] which shows that the excitation of 1849, unlike that of 2537, is mainly directly from the normal state.

IONIZATION OF THE 6^3P STATES

The work of Klarfeld[35] has indicated that the cross sections for the ionization of the 6^3P states are an order of magnitude higher than that for ionization of 6^1S_0. Unpublished calculations made on this basis, and observations in this laboratory to be reported later, have indicated that practically all the ionization in the present lamp comes from 6^3P. The importance of such transfers for the present picture can be estimated from observed positive ion current to the wall, $\sim 50\mu$a cm^{-2}, or $\sim 500\mu$a per cm of length. This amounts to 3×10^{15} ions sec.$^{-1}$ cm^{-1}, a relatively negligible effect. It is to be noted that on account of the relative nearness of 6^3P_2 to ionization and of its relatively high population, most of the ionization will come from this state. Direct experimental evidence of this has been found in this laboratory and will be published elsewhere.

DISCUSSION

The various transitions shown in Fig. 4 have been summarized in Table IV, with a view to showing the degree of balance attained for each state. As for 6^3P_0 it is to be remembered that transitions to and from 6^3P_1 were chosen partly with a balance in view. In the other cases, no such arbitrary adjustments were made, the balances depending on 6^3P populations based on measured 2537 output; 2537 diffusion theory and relative absorption measurements; measured electron temperature and concentration and published experimental and theoretical excitation cross section curves. The degree of balance attained among all the processes is considered to be better than the accuracy of some of the data warrant. But the self-consistency shown is believed to be good evidence that the over-all picture cannot be far wrong. Possibly it can be used as evidence for the essential correctness of some of the data used, e.g., the more important cross sections.

TABLE IV. Summary of the transition rates represented in Fig. 4, showing the degree of balance attained for each of the 6^3P states and 7^3S_1 states.

State 1	State 2	Transitions* arriving at state 1 from state 2	Transitions* leaving state 1 for state 2	Net gain %
6^3P_2	6^1S_0	4.5	0.95	
	6^3P_1	1.09	4.83	
	6^3P_0	3.0	3.0	
	7^3S_1	$0.17L$**	0.16	
	6^3D	$0.06L$	0.8	
	Total	8.82	9.02	−1.1
6^3P_1	6^1S_0	3.01	{0.14 {$8.30L$	
	6^3P_2	4.83	1.09	
	6^3P_0	1.36	0.31	
	7^3S_1	$0.23L$	0.024	
	6^3D	$0.05L$	0.011	
	Total	9.48	8.90	6.5
6^3P_0	6^1S_0	1.33	0.28	
	6^3P_2	3.0	3.0	
	6^3P_1	0.31	1.36	
	7^3S_1	$0.09L$	0.036	
	6^3D	$0.01L$	0.017	
	Total	4.73	4.69	1.0
7^3S_1	6^1S_0	0.09	—	
	6^3P_2	0.158	$0.17L$	
	6^3P_1	0.024	$0.23L$	
	6^3P_0	0.036	$0.09L$	
	Total	0.31	0.49	−37.0

* In units of 3×10^{16} sec.$^{-1}$ cm^{-1} (arrows in Fig. 4).
** L refers to measured radiation.

Panevkin,[18] studying a decaying plasma, initially essentially like the present discharge, observed, by absorption measurements, decay times for $6^3P_{2,1,0}$ of 4×10^{-5}, 1×10^{-5}, and 6×10^{-5} sec. respectively. The value for 6^3P_1 namely 1×10^{-5} sec. agrees fairly well with that predicted here (0.76×10^{-5} sec.), considering that in Panevkin's case this state is being replenished from $6^3P_{2,0}$ by electronic collisions (the electrons disappearing by diffusion and recombination relatively slowly on this time scale). Indeed on the present picture it would have been expected that after an initial short period of adjustment this replenishment would have caused 6^3P_1 to decay along with the metastables. It is believed that this may have been the case, but was not shown up in Panevkin's experiments because of the difficulty of making the absorption measurements on 6^3P_1, at appreciable times after cut-off, with the necessary accuracy.

The present scheme accounts rather well for the initial rates of decay of the 6^3P states, especially $6^3P_{2,1}$, found by Panevkin.

According to Fig. 4 about one third of the excitation of 2537 comes directly from 6^1S_0 and two thirds stepwise via the metastables. The same conclusion was arrived at by Panevkin[18] and earlier by Butaeva and

[34] F. Butaeva and V. Fabrikant, J. Tech. Phys. U.S.S.R. 18, 1127 (1948).
[35] B. Klarfeld, Tech. Phys. U.S.S.R. 5, 913 (1939); Comptes Rendus U.S.S.R. 24, 251 (1939).

Fabrikant.[36] These authors do not give the details of their calculations. Since as noted above Panevkin has used 6^3P populations which are one fourth to one third as great as those found here, it is to be inferred that he has used quite different values of T_e and n_e or else that his cross sections for transfers among the 6^3P states were several-fold larger than those used here. However, he quotes the same authority for these as that used here.[22]

Evidence for the essential correctness of the present scheme is found in the fact that it can be used along with the observed change in voltage to estimate with some accuracy the effect of changing Hg density and changing current on the 2537 output, and on the electron temperature and concentration.

The method consists in finding a new balance, corresponding to Fig. 4 which will satisfy the new conditions; it may be illustrated as follows. Suppose it is desired to find the effect of increasing by a certain factor the concentration n_a of Hg atoms, the current being kept constant. A consideration of the 2537 diffusion problem shows that the time of imprisonment T will increase roughly as $n_a^{1.17}$. The average concentration n_1 of 6^3P_1 atoms will as before be related to the 2537 output R quanta per cm length of the tube per sec. by the relation

$$\pi a^2 n_1 = RT.$$

The concentration n_1 will increase with T (although R may or may not, depending on the initial n_a) and this will allow the populations n_2 and n_0 of $6^3P_{2,0}$ to increase. Larger populations of the 6^3P states will, other things being equal, give a larger excitation of the visible and near ultraviolet lines and, more importantly for the discharge, a larger rate of ion production. The last result causes the observed drop in voltage and an increase in n_e (obtained with the use of the mobility equation), and a decrease in T_e.

It is now possible to find, largely by trial and error, a new T_e and a new corresponding set of 6^3P populations which will balance the excitation and de-excitation processes, and at the same time yield the needed rate of ion production (assumed proportional to n_e). It has been possible in this way, without too much labor, to predict a maximum in the 2537 output at around 40°C (6.1μ Hg pressure) as is observed.[37,38]

When the current (alone) is increased, T will not change but all processes depending on n_2 will increase. Most important is the resultant increase in n_1, which allows n_2 and n_0 to increase slightly, other things being equal. This increase in the 6^3P populations allows more efficient ion production, but it also gives a greater relative excitation of the visible and near ultraviolet lines at the expense of 2537. It has been found possible by the above method to find a new T_e, and n_e (corresponding with the observed drop in voltage) which will yield a new set of 6^3P populations, balancing all excitation and de-excitation processes and at the same time yield the needed ionization. The result predicts a decrease in the production efficiency of 2537 in good accord with the observed.

Butaeva and Fabrikant[34] have given an interesting discussion of the variation of 2537 output with Hg pressure and current but this method cannot suffice for the present case because it does not take into account the variation in T_e.

In the scheme of Fig. 4 the populations of the 6^3P states are held down to the observed small fraction of the Boltzmann populations by the "leakage" of 2537. It was seen above that about 84 percent of the total excitation of $6^3P_{2,1,0}$ is radiated as 2537, the remainder or 16 percent being returned to normal by collisions of the second kind. As the current (alone) increases, the latter percentage will increase, and when it has become all important and the 2537 leakage (and efficiency of production) negligible, the true Boltzmann populations will have been reached. In this process 6^3P_1 will increase relatively faster than 6^3P_2 and 6^3P_0, but due to falling T_e it may be that the absolute populations of $6^3P_{2,0}$ will decrease slightly as Panevkin[21] has shown (see also Fig. 1).

The writer is indebted to many of his colleagues for helpful discussions and particularly to Dr. C. G. Found, Dr. N. L. Oleson (now of the Postgraduate School, U. S. Naval Academy, Annapolis, Md.), Dr. B. T. Barnes, and Miss M. A. Easley.

[36] F. Butaeva and V. Fabrikant, Bull. Acad. Sci. U.S.S.R. Ser. Phys. 9, 230 (1945).

[37] W. E. Forsythe and E. Q. Adams, reference 3, p. 154ff and references therein.

[38] General Electric Lamp Bulletin LD-1, 59 (1946).

Vibrational Relaxation Times in Gases*

WAYLAND GRIFFITH†
Department of Engineering Sciences and Applied Physics, Harvard University, Cambridge, Massachusetts
(Received July 17, 1950)

Relaxation times for seventeen gases have been measured using a steady flow method, the velocity field being found numerically by Southwell's Relaxation Method. Where a comparison is possible the results are found to be in good agreement with sound dispersion data. The present state of knowledge about relaxation effects is discussed and a few instances of their possible appearance are cited.

INTRODUCTION

RECENT advances in the study of high speed flows have aroused new interest in the kinetic processes of fluid mechanics. In very strong shock waves, for example, conditions required by the Rankine-Hugoniot relations may be attained only outside a region considerably thicker than that usually estimated assuming internal energy equilibrium. Historically, the fact that internal energy states of molecules fail to keep up with sufficiently fast changes in their environment was first observed experimentally in 1925. Pierce[1] was able to measure dispersion of sound which could be explained only by ascribing a relaxation time to the adjustment of vibrational energy in addition to the effects of heat conduction and viscosity. Since then measurements have been made on dispersion and absorption of sound in several pure gases and many mixtures, and suitable theories advanced for interpreting the results.[2]

In 1942 Kantrowitz[3] proposed a very clever method for finding relaxation times using a steady flow of a gas. When the total pressure in a jet is measured with a pitot tube the value will be equal to the pressure in the reservoir from which the jet issues, provided the effects of viscosity and heat conduction are negligible, and provided all changes in the flow are slow compared to the rates of energy exchange between molecules. The former are usually limited to the boundary of the flow next to the walls, outside this region their effect is small. If the gas is now imagined to stop at the nose of a pitot tube quickly compared to the time for adjustment of vibrational energy, say, the compression will take place as though the gas had no vibrational energy, the specific heat ratio γ will be altered, and the resulting pressure will be different. The relaxation time enters the calculation through the time dimension involved in the stream velocity and pitot tube diameter. Kantrowitz verified the reliability of this method with measurements on carbon dioxide.

The purpose of the present investigation is threefold; (a), to measure vibrational relaxation times for several gases, (b), to extend confidence in the results of this method by using careful techniques and an exact solution of the flow pattern, and (c), to find as nearly as possible the limits of the method, for example in detecting very fast adjustments.

In the following section a very brief presentation of the thermodynamics appropriate to these experiments will be given, along with a physical explanation which it is hoped will aid in understanding the kinetic processes involved. Also, "relaxation time" will be defined.

THERMODYNAMICS FOR A GAS WITH LAGGING INTERNAL ENERGY

In addition to the usual thermodynamic variables for a gas it is appropriate to define a new one which describes the energy of vibration in the molecules. There are fairly good reasons to suppose that the various modes are in equilibrium among themselves even though they exchange energy slowly with the external motions of translation and rotation. In the first place there is always a certain amount of anharmonic coupling. Secondly, the lifetime is found experimentally to extend over hundreds or thousands of collisions. Although no direct evidence is available for it, a mechanism by which vibrational energy is redistributed within a molecule during a collision would have ample chance to operate. Accordingly T_{vib} will be defined as the temperature in the Boltzmann formula corresponding to the population distribution among vibrational states.

Three variables are therefore needed to specify a thermodynamic state for the gas, pressure and temperature say, and T_{vib}. The specific heat will be regarded as arising from two sources, vibrational motion, and all the rest. Following the nomenclature of Kantrowitz $c_{\text{vib}} \equiv dE_{\text{vib}}/dT_{\text{vib}}$, where E_{vib} is the energy in vibrational motion. C_p' will refer to the specific heat at constant pressure not counting vibration. Thus the usual specific heat $C_p = C_p' + C_{\text{vib}}$ when $T = T_{\text{vib}}$. Similarly C_v and C_v' will be used for constant volume specific heats. The Ideal Gas equation of state is assumed throughout.

The entropy change in going from one state to another is $\Delta S = \int dQ/T$, where any reversible path may be used to evaluate ΔS. By considering such a reversible change between two states which are allowed to approach one another indefinitely, one easily arrives at a

* Doctoral Thesis, Harvard University, 1949.
† Now at Palmer Physical Laboratory, Princeton, New Jersey.
[1] G. W. Pierce, Proc. Am. Acad. Sci. 60, 271 (1925).
[2] An excellent review on sound dispersion and a bibliography is given by W. T. Richards, Rev. Mod. Phys. 11, 36 (1939).
[3] A. Kantrowitz, J. Chem. Phys. 10, 145 (1942).

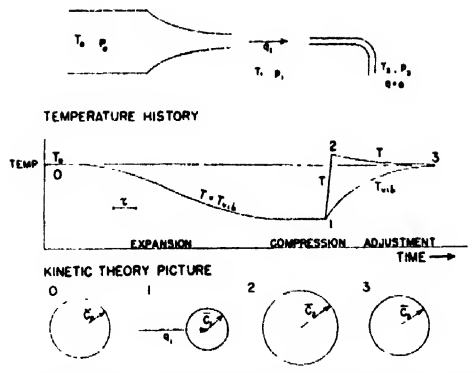


FIG. 1. Schematic drawing of the experiment, temperature history of the gas, and the variation in velocities according to kinetic theory.

formula for the entropy change as

$$dS=C_p'dT/T+C_{\text{vib}}dT_{\text{vib}}/T_{\text{vib}}-Rdp/p. \quad (1)$$

A familiar form of the Second Law results when T becomes equal to T_{vib}. We see that this is strictly applicable only when changes in temperature are slow compared to internal adjustments.

In many flow problems involving gases it is possible to neglect the viscosity, heat conductivity, and external forces in comparison to the other terms, and to assume constant specific heats without introducing serious errors. The equations for a compressible fluid with possible separate translational and vibrational temperatures are then

Continuity $$\partial\rho/\partial t+\partial\rho q_i/\partial x_i=0 \quad (2)$$

Momentum $$\rho\frac{dq_i}{dt}=-\partial p/\partial x_i \quad (3\text{–}5)$$

Energy $$\frac{\partial}{\partial x_i}\rho q_i\left(C_{\text{vib}}T_{\text{vib}}+C_v'T+\frac{p}{\rho}+\frac{q_j^2}{2}\right)$$

$$=-\frac{\partial}{\partial t}\rho\left(C_{\text{vib}}T_{\text{vib}}+C_v'T+\frac{q_j^2}{2}\right). \quad (6)$$

Eliminating the velocity from the energy and momentum equations gives

$$d/dt(C_{\text{vib}}T_{\text{vib}}+C_p'T)-(RT/p)dp/dt=0. \quad (7)$$

Combining this with (1) gives the entropy as a function of T and T_{vib};

$$dS=(1/T_{\text{vib}}-1/T)dC_{\text{vib}}T_{\text{vib}}. \quad (8)$$

This is what one might write down intuitively by considering the transfer of energy $dC_{\text{vib}}T_{\text{vib}}$ from the temperature T to T_{vib}.

A consideration of the actual collisions is required in order to find an expression for the rate at which the above exchange will take place. Using simplified models Bethe and Teller derived formulas predicting the decay function for adjustment and its dependence on pressure, vibration frequency, and temperature. Details in the steps are not included here but may be found in the papers by Bethe and Teller,[4] Landau,[5] and Landau and Teller.[6]

The variation of T_{vib} is found to be

$$dC_{\text{vib}}T_{\text{vib}}/dt=C_{\text{vib}}(T-T_{\text{vib}})/\tau, \quad (9)$$

where the relaxation time has the following functional dependence;

$$\tau\approx p^{-1}(\nu s)^{-\frac{1}{3}}\left(\frac{m}{T}\right)^{-\frac{1}{2}}\exp\left\{(\nu s)^{\frac{1}{3}}\left(\frac{m}{T}\right)^{\frac{1}{2}}\right\}. \quad (10)$$

The new quantities are ν-vibration frequency, s-molecular radius, and m-molecular mass. When changes in the temperature are small τ may be considered constant and after a time τ the temperature difference diminishes to $1/e$, or 0.368, of its original value.

With this information let us study the case mentioned earlier of a slow expansion through a nozzle to a velocity q_1 and a fast compression in front of a pitot tube. From the equations of momentum and energy

$$d/dt(C_{\text{vib}}T_{\text{vib}}+C_p'T+q^2/2)=0. \quad (11)$$

By this and Eq. (1) the ratio of reservoir pressure to pitot tube pressure will be

$$p_0/p_3=\exp\{\Delta S/R\}. \quad (12)$$

The entropy change is found from (8) and (9), the increase occurring while T and T_{vib} converge upon the stagnation value T_0. (See Fig. 1.)

$$\Delta S=\int_{T_1}^{T_0}\left(\frac{1}{T_{\text{vib}}}-\frac{1}{T}\right)dC_{\text{vib}}T_{\text{vib}}$$

$$=\int_{T_1}^{T_0}\left(\frac{1}{T_{\text{vib}}}-\frac{C_p'}{C_pT_0-C_{\text{vib}}T_{\text{vib}}}\right)dC_{\text{vib}}T_{\text{vib}}.$$

This finally gives,

$$\frac{p_0}{p_3}=\left(\frac{T_0}{T_1}\right)^{C_p/R}\left(\frac{C_p'T_0}{C_pT_0-C_{\text{vib}}T_1}\right)^{C_p'/R}. \quad (13)$$

The pressure at state 2 may also be found quite easily. Equations (9) and (1) give,

$$\frac{C_p'}{R}\frac{dT}{T}-\frac{dp}{p}+\frac{C_{\text{vib}}dt}{\tau RT}(T_{\text{vib}}-T)=0. \quad (14)$$

If τ is very long compared to other times in which changes take place in isentropic flow the last term

[4] H. A. Bethe and E. Teller, Report X-117, Ball. Res. Lab., Aberdeen Proving Ground.
[5] L. Landau, Physik. Zeits. Sowjetunion 1, 89 (1934).
[6] L. Landau and E. Teller, Physik. Zeits. Sowjetunion 10, 34 (1936).

vanishes and

$$p/p_0=(T/T_0)^{C_p'/R} \tag{15}$$

that is, the gas acts as though it had only the specific heat C_p'. For τ very short on the other hand, $T=T_{vib}$ and

$$\frac{p}{p_0}=\left(\frac{T}{T_0}\right)^{C_p/R} \tag{16}$$

in the usual fashion. From (11)

$$C_p(T_0-T_1)=\frac{q_1^2}{2}=C_p'(T_2-T_1). \tag{17}$$

Solving we find

$$\frac{p_0}{p_2}=\left(\frac{T_0}{T_1}\right)^{C_p/R}\left(\frac{C_p'T_0}{C_pT_0-C_{vib}T_1}\right)^{C_p'/R}. \tag{18}$$

The surprising result is that p_2 and p_3 are equal so that the adjustment of internal energy takes place at constant pressure. In his thesis* the author shows that $p_0>p_2$, which is consistent with the increased entropy.

Figure 1 shows a schematic drawing of the experiment, the time-history of the process, and a sort of kinetic theory picture of the situation at various stages of the flow. The acceleration from 0 to 1 is a conversion of random molecular motion with average velocity $\bar{c}_0$ into streaming motion of flow velocity q_1 and smaller average velocity $\bar{c}_1$. The compression from 1 to 2 returns all the stream energy into random motion without reexciting the vibrations immediately. The average velocity is higher than in the reservoir. The pressure is less, however, as the density is sufficiently lower to compensate for the higher temperature. The adjustment of internal energy takes place at constant pressure, decreasing temperature, and increasing density. The average molecular velocity returns to its original value but the gas is less dense.

With this qualitative idea of the effect in mind we can now develop the methods used for determining relaxation times from pressure measurements. Again we shall use a symbol introduced by Kantrowitz as a measure of the disequilibrium between vibration and translation,

$$\epsilon\equiv c_{vib}(T_{vib}-T). \tag{19}$$

In flows where changes in T are small compared to T itself (8) may be written

$$ds=-\frac{\epsilon dt}{\tau}\frac{T-T_{vib}}{TT_{vib}}\cong\frac{\epsilon^2 dt}{\tau C_{vib}\bar{T}^2}. \tag{20}$$

$\bar{T}^2$ is taken as the average of TT_{vib} over the whole process. From Fig. 1 a good value is seen to be $\bar{T}=\frac{1}{2}(T_0+T_1)$. The specific heats and relaxation time will be taken at this temperature as well.

When the difference $p_0-p_3=\Delta h$ is small compared to p_0,

$$\Delta S/R\cong\Delta h/p_0 \tag{21}$$

and

$$\Delta h/p_0=\int\frac{\epsilon^2 dt}{\tau C_{vib}\bar{T}^2} \tag{22}$$

where ϵ depends on the relaxation time and the history of the flow. The integral is to be taken over that part of the flow for which Δh is desired. An equation for ϵ is found by eliminating the temperature from (9), (11), and (19),

$$d\epsilon/dt+C_p\epsilon/C_p'\tau=C_{vib}/C_p'd(q^2/2)/dt \tag{23}$$

the integral being

$$\epsilon(t_1)=\exp\left(-\int^{t_1}\frac{C_p dt}{\tau C_p'}\right)\int^{t_1}\times\exp\left(\int^{t_1}\frac{C_p dt}{\tau C_p'}\right)\frac{C_{vib}}{C_p'}\frac{d(q^2/2)}{dt}dt \tag{24}$$

This completes the thermodynamics, for a knowledge of the velocity history of the flow and the specific heats now enable one to relate pressure differences to relaxation times.

DESIGN OF THE EQUIPMENT

A review of the literature on sound dispersion indicates that relaxation times are of the order of five microseconds. Generally diatomic molecules have longer times, polyatomic gases shorter times. Accordingly a steady flow is desired in which a gas expands slowly compared to this time and stops in about five microseconds.

Because the analytical flow solutions known are limited in number and sometimes difficult to reproduce experimentally, it was decided to use a fairly simple arrangement with axial symmetry and solve the problem numerically using R. V. Southwell's Relaxation

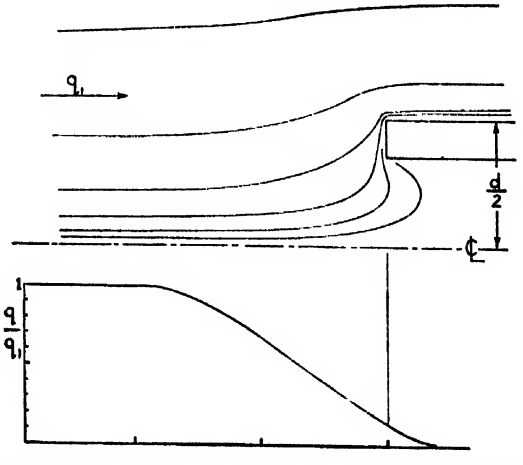


FIG. 2. Streamlines for flow past an open-ended cylinder computed numerically and the velocity along the central streamline.

Method.[7] It rests upon the accuracy with which a difference equation can be made to approximate a differential equation. An excellent discussion of the method may be found in a paper by Emmons.[8] Incompressible flow was assumed, the effect of compressibility can be found from the velocity dependence of the observations.

The solution for a plane parallel jet impinging on a flat plate was obtained first since the simplicity of this arrangement was quite attractive. The required nozzle diameter was only 0.024 in., though, so this scheme was abandoned in favor of a pitot tube to obtain sufficiently fast stopping times.

A square-ended cylinder was chosen for the shape of the tubes and the Relaxation Method used to locate streamlines. First the streamline at 2.25 d was assumed to be undisturbed; when this was moved out to 2.50 d values of the stream function were changed by less than one percent. A hollow center of diameter 0.620 d was

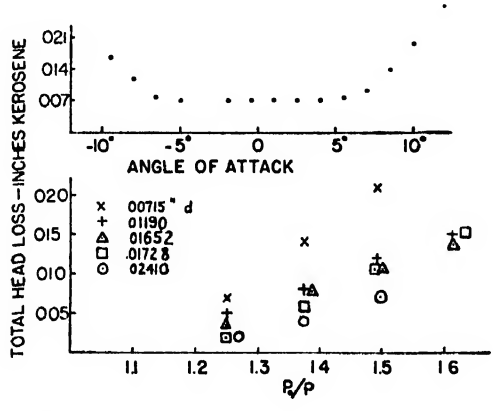


FIG. 3. The effect of angle of attack on total head loss and minimum losses for five tubes at various ratios of reservoir to stream pressure.

then included in the calculation. This made the deceleration of the flow considerably less abrupt. The streamlines and a velocity-time plot are shown in Fig. 2. For a jet velocity of 400′/sec. the diameter should then be about $400\times12\times5\times10^{-6}\times2/3=0.016$ in. It is interesting that the disturbance extends only one diameter upstream.

Glass tubes were found to be most satisfactory, being both easy to clean and relatively easy to make. With sufficient practice and patience square ends on tubes down to 0.005 in. could be ground with a fine Water of Ayr stone. Diameters of the tubes actually used in various phases of the investigation were as follows: 0.0241 in., 0.01728 in., 0.01652 in., 0.01190 in., and 0.00715 in. The respective i.d./o.d. ratios were: 0.610, 0.622, 0.590, 0.604, and 0.622.

A good design for the nozzle is important since we should like to measure only the effect due to fast stopping. Equation (23) shows that the best possible type of acceleration is one in which dq^2/dt is constant. This is the same as dT/dt constant, chosen by Kantrowitz for another reason. Two nozzles one-half inch long were made. The exit diameters were 0.070 in. and 0.040 in. The flow at 400′/sec. is 1.1 cfm through the larger nozzle and 0.36 cfm through the smaller one. A correction for the wall boundary layer was made, similar to that described in Kantrowitz's paper.

The reservoir was made from a two-foot section of three inch iron pipe. The nozzle was mounted in one end, the other having a connection for the gas supply line and a blowout diaphragm. Inside the chamber a 100-mesh screen and straightening vanes were placed twelve inches from the nozzle. The reservoir was mounted horizontally and had a thermocouple and pressure tap for measuring the stagnation conditions. Arrangements for a few experiments at elevated temperatures were made by placing a coil gas heater in the supply line and winding the reservoir with Nichrome wire and pipe insulation.

A micromanometer was devised to measure small pressure differences very accurately and quickly. A short length of glass tubing with a fine cross-hair was mounted at an angle on a frame which travelled vertically. Also mounted on the frame was an Ames dial which read distances above an arbitrary level to an accuracy of 0.0005 in. With a mirror fixed behind the cross-hair one could move the frame to align the meniscus and hairline within 0.003 in. For a slope of 1:4 in the tube, the height of the meniscus could be measured to 0.001 in. Since the meniscus is always in the same part of the tube the usual difficulties of varying wall diameter and incomplete wetting were greatly reduced. With kerosene in the manometer a change in pressure of 2×10^{-6} atmos. was detectable.

A special holder was made for the tubes which allowed independent adjustments to be made in axial and transverse positions and in direction. Once the tip of the tube was properly located in the stream the angle of attack could be changed to align it correctly. The tubing used was Seran, a transparent plastic. Reservoir pressure was measured with a mercury U-tube. All of the pressure lines were kept at one level as nearly as possible to reduce errors due to buoyancy when gases with densities quite different from air were tested. A small blower and stovepipe were used to draw off the more toxic gases.

EXPERIMENTAL METHODS AND DETERMINATION OF RESULTS

The gases used were obtained in steel cylinders from commercial supply houses. The purest form available was chosen in each case since work on sound dispersion indicated that certain impurities, especially water vapor, could cause large variations in the measured results. Preliminary runs with compressed air showed

[7] R. V. Southwell, *Relaxation Methods in Engineering Science* (Clarendon Press, Oxford, 1940).
[8] H. W. Emmons, Q. App. Math. 2, 3, 173 (1944).

that readings could be taken once a minute. Nitrogen was chosen as the best available gas for taking minimum readings because of its very small vibrational heat capacity. The effect of the angle of attack is shown in Fig. 3. The range of minimum loss extends ±5°, thus making proper alignment fairly easy. Consistent readings were found for axial positions between one and two nozzle diameters and over three-fourths of the stream. Also shown in Fig. 3 are the minimum pressure losses at various pressure ratios using the 0.040 in. nozzle. The readings were very nearly the same for the 0.070 in. nozzle. The dependence on tube size cannot necessarily be interpreted as a relaxation effect, for the smaller tubes may be less perfect, and some loss due to viscosity might also be expected. Important to the main work is the conclusion that the losses due to other than relaxation effects can be reduced to such a small value. Before each experimental run readings with nitrogen were taken and subtracted from the observed data. A qualitative idea of relative relaxation times may be formed from Fig. 4 which shows the measured pressure loss for several of the gases tested.

The method for calculating relaxation times from these data will now be developed. Equation (23) gives

$$(d\epsilon/dt)dt+(C_p\epsilon/C_p')dt/\tau = (C_{vib}/C_p')\left(d\frac{q^2}{2}\Big/dt\right)dt. \quad (25)$$

As the time required for stopping becomes shorter the second term will become small compared to the other terms, and in the limit, when $\int \epsilon dt/\tau$ becomes negligible in comparison to the other two terms, the change in ϵ is

$$\Delta\epsilon=(C_{vib}/C_p')\Delta q^2/2.$$

In the case being considered ϵ was originally zero and the velocity was q_1. A sudden stopping then produces a value of

$$\epsilon_0=(C_{vib}/C_p')q_1^2/2.$$

With $q=0$, (25) gives for the approach to equilibrium $d\epsilon/dt=-C_p\epsilon/C_p'\tau$. The pressure change for an instantaneous compression is accordingly,

$$\Delta h_{ic}/p_0=\int_{\epsilon_0}^{0}\frac{\epsilon^2 dt}{\tau RC_{vib}\bar{T}^2}=\int_{\epsilon_0}^{0}\frac{C_p'\tau\epsilon d\epsilon}{C_p\tau RC_{vib}\bar{T}^2} =\frac{C_p'}{2C_pC_{vib}R\bar{T}^2}\left(\frac{C_{vib}q_1^2}{2C_p'}\right)^2. \quad (26)$$

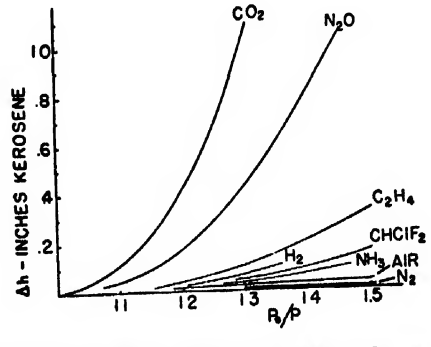


FIG. 4. Observed pressure losses for several of the gases tested.

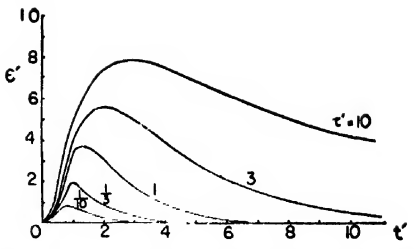


FIG. 5. The parameter ϵ' measures the disequilibrium of vibrational energy. For instantaneous stopping it would initially have the value 1; plotted are the variations in ϵ' for five values of the dimensionless relaxation time τ'.

It is more convenient at this point to substitute a set of dimensionless variables than to continue with the physical variables. Using Kantrowitz' notation again define the following dimensionless variables,

$$\tau'=\frac{C_p'}{C_p}\frac{q_1\tau}{d} \qquad t'=\frac{q_1 t}{d}$$
$$\epsilon'=\frac{\epsilon}{(C_{vib}/C_p')(q_1^2/2)} \qquad q'=q/q_1. \quad (27)$$

Using (27) and (26) in Eqs. (24) and (22) we find

$$\epsilon'=\exp\left(-\frac{1}{\tau'}\int dt'\right)\int\exp\left(\frac{1}{\tau'}\int dt'\right)\frac{dq'^2}{dt'}dt' \quad (28)$$

and

$$\frac{\Delta h}{\Delta h_{ic}}=\frac{2}{\tau'}\int\epsilon'^2 dt' \quad (29)$$

Thus we see that $\Delta h/\Delta h_{ic}$ depends only on τ' for a given flow pattern. This is the great simplification obtained by running an experiment in which only small temperature changes occur.

The integration of (28) to find ϵ' was done numerically for several values of τ'. Curves of ϵ' corresponding to these are given in Fig. 5. The mathematical difficulty of infinite stopping time on the central streamline was avoided by abandoning the integration when further contributions from the integrand became negligible. Physically the discrete nature of the gas explains the paradox.

The integration for $\Delta h/\Delta h_{ic}$ was found by counting squares in a graph of ϵ'^2 *vs.* t'. The resulting values are

τ'	1/10	1/3	1	3	10
$\Delta h/\Delta h_{ic}$	0.0828	0.226	0.480	0.723	0.908

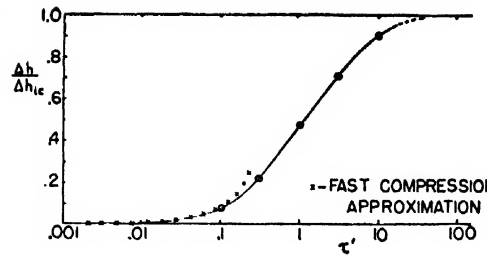


FIG. 6. Graph used to find dimensionless relaxation time τ' from observed pressure loss Δh.

For very short relaxation times an approximation method used by Kantrowitz was also applied here, the result being $\Delta h/\Delta h_{ic}=0.867\tau'$. Figure 6 shows that this is valid for $\tau'<0.08$.

The temperature T_1 was found from the pressure ratio p_1/p_0 and Eq. (16), the velocity q_1 from (17). Thus Δh_{ic} could be computed and τ' read from the graph.

EXPERIMENTAL RESULTS

The results for N_2O shown in Fig. 7 give an idea of the consistency of the method and show that no velocity dependence up to $M=0.8$ is observable. A similar test on CO_2 confirmed this result. Apparently the faster deceleration in the compression is just compensated for by the gain in collision rate. Relaxation times for the seventeen gases tested are given in Table I. The temperature dependence of τ shown in Fig. 8 was found from a few measurements; the accuracy was somewhat reduced due to difficulties in keeping the reservoir temperature equal to the gas temperature.

Hydrogen has no appreciable vibrational energy at room temperature so the observed pressure loss was ascribed to rotational lag and the same formulation assumed applicable. An upper limit for rotation of N_2 was also computed in this way. Only an upper limit could be found for four gases which showed no pressure loss greater than N_2. Two brands of CO_2 were tested, commercial CO_2 with a few percent water vapor, and a special purified lot containing 0.044 percent water vapor. This produced a large change in the result and it is felt that the time for the pure gas is somewhat longer still. This conclusion is borne out by a comparison with sound dispersion data, the most recent value for carefully purified CO_2 being about a third larger. Other impurities in other gases are not nearly as effective as this combination so the results may be considered fairly reliable. As a possible check for future work the suppliers' data on impurities are included in Table I. Also included are values from the literature on all the pure gases tested. Of the six in common beside CO_2 agreement is seen to be very good for H_2, N_2O, and C_2H_4, while agreement for SO_2, CH_4, and NH_3 extends only to the order of magnitude.

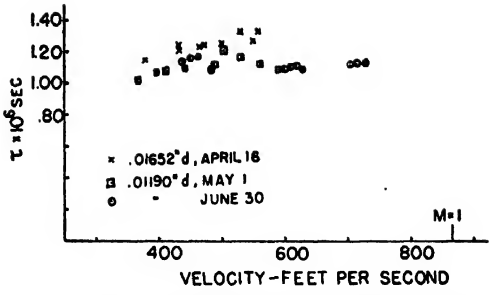


FIG. 7. Data on N_2O indicating accuracy of readings and absence of compressibility effect on results.

DISCUSSION

With the information on pure gases available from these experiments and the data already reported in the literature several conclusions may be drawn about the collisions. Relaxation times decrease for increasing temperature, supporting the theory that the fast collisions are most important. The higher the vibration frequency the longer the relaxation time; $CH_4>CH_3Cl>CHF_2Cl$; $CH_4>C_2H_6$; $C_2H_4>C_3H_6$. This indicates that the idea of resonant coupling used in deriving (10) is fairly reasonable. Since progression along the methane and ethylene series seems to give shorter relaxation times it is likely that all complicated molecules reach equilibrium quickly. Few fruitful measurements have been made on diatomic molecules, presumably because their small vibrational energy produces an exceedingly small effect. A further difficulty arises from the long times indicated; an impurity that is more effective than the molecule itself in collisions will control the adjustment.

This brings up the next important consideration, the effect of mixtures. Early experimenters were not aware of the great influence which foreign molecules with chemical affinity have. This accounts in large part for the dismaying array of values in the literature, and for the remarkable difference in the two brands of CO_2 tested here. The shift in the absorption maximum has been found linear for small amounts of impurities,[9] indicating that bimolecular collisions are responsible for energy exchange. A single exception is the case of O_2, where a quadratic dependence on water vapor content

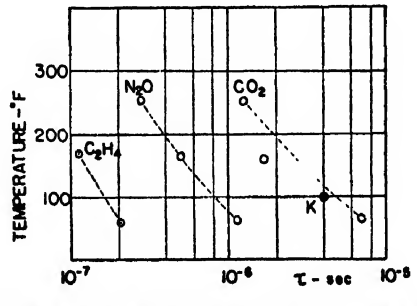


FIG. 8. Dependence of relaxation time on temperature. The point marked K is from Kantrowitz' work.

[9] V. O. Knudsen and E. F. Fricke, J. Acous. Soc. Am. 12, 255 (1940).

points to triple collisions as the mechanism.[10] In general unlike molecules exchange energy more readily than like molecules. Since the monatomic gases and N_2 are exceptions[11] one may conclude that chemical affinity is important. The small mass and high velocity of H_2 is significant since it is more effective than D_2. To explain the orders of magnitude the moments of inertia and dipole strength must also be considered.[11,12] Eucken[13] discusses the problem from the point of view that vibrational excitation is preliminary to chemical reaction; if colliding molecules are known to react exchange of vibrational energy is likely.

Limitations of the flow method of measurement can now be outlined fairly well. The size of nozzles and pitot tubes used here represent the lower practical limit of utility and construction. A sharp-nosed tube might produce half the stopping time, but would be harder to make. Operating the entire experiment at a lower pressure lengthens the adjustment period, but reduces the sensitivity even more. For testing compounds with small vapor pressures this is necessary anyway. For diatomic molecules higher pressures would give greater sensitivity, the disadvantage being the larger gas flow. The inherent inability of this method to detect more than one rate of adjustment does not seem to be serious, since no bona fide cases of this sort have been discovered in sound dispersion.

CONCLUSIONS

The results of this investigation extend considerably the available data on pure gases. Molecular energy exchange is such a complicated process that it is unfortunately still impossible to predict a relaxation time for an untested gas or mixture. Certain limits, however, may now be set on conditions where relaxation effects will appear. Calculations show that the thickness of a shock wave in CO_2 is about a tenth of an inch; at $M=2$ the initial value of $T-T_{vib}$ is 350°F so quite a strong effect in this narrow region might be expected. Experiments to look for this using a shock tube are being planned. In the usual aerodynamic problems relaxation effects are completely negligible as has always been assumed. For extreme cases of heat and velocity as are met in rocket and meteor flight studies relaxation effects are doubtless significant, but until the large-scale effects are better understood this will not be considered.

The flow method is seen to be a very satisfactory technique for measuring vibrational relaxation times, and may be used to study many other rate processes, for example dissociation and recombination.

The author would like thank Professor Howard Emmons for many valuable discussions and continued interest during the course of this research.

TABLE I.

Gas	Temp. °F	Relaxation time μsec.	Gas purity and foreign gas Present investigation
Hydrogen	58	0.021 (rot.)	99.8%, 0.2% H_2O
Nitrogen	60	<0.002 (rot.)	
Nitrous oxide	65	1.12	98–99%, 1–2% N_2
Carbon dioxide	65	6.95	99.956%, 0.044% H_2O
Carbon dioxide (comm.)	65	2.02	n% H_2O
Sulfur dioxide	68	0.37	99.988%, 0.002% H_2O
Ammonia	63	0.12	99.5%, 0.005% H_2O
Methane	60	0.48	99%, 1% C_2H_6, N_2, CO_2
Methyl chloride	72	0.202	99.5%, <0.008% H_2O
Freon 22	70	0.10	97%, <0.005% H_2O
Freon 12	67	0.090	95%, <0.001% H_2O
Ethane	57	0.003	95%, 5% C_2H_4, C_3H_8, C_3H_6
Ethylene	60	0.207	99.5%, 0.5% air
Propylene	70	<0.006	99%
Propane	70	<0.007	99%
Ethylene oxide	86	1.23	99.8%
Butane	70	<0.018	99%
Butadiene	70	<0.013	99.77%, 0.23% C_4H_{10}
Results reported in the literature:			
Oxygen	70	1000	Reference [a]
Carbon monoxide	1800	10	[b]
Chlorine	70	18	[c]
Nitric oxide	60	0.8 (el. excit.)	[d]
Hydrogen	77	0.018 (rot.)	[e]
	60	0.02 (rot.)	[f]
Deuterium	60	0.015 (rot.)	[f]
Carbon dioxide	70	10.8	[f]
	70	10.26	[g]
Nitrous oxide	67	0.92	[e]
	70	1.44	[g]
Carbonyl sulfide	70	0.86	[g]
	70	1.67	[e]
Carbon disulfide	70	0.70	[g]
Sulfur dioxide	70	0.181	[g]
Water vapor	415	0.037	[h]
Ammonia	70	0.4	[i]
Ethylene	70	0.238	[j]
Methane	230	0.84	[e]

[a] See reference 12.
[b] G. G. Sherratt and E. Griffiths, Proc. Roy. Soc. London **147A**, 292 (1934).
[c] See reference 13.
[d] See reference 10.
[e] E. S. Stewart, Phys. Rev. **69**, 632 (1946).
[f] A. van Itterbeek and R. Vermaelen, Physica **9**, 345 (1942).
[g] E. F. Fricke, J. Acous. Soc. Am. **12**, 245 (1940).
[h] A. Kantrowitz and P. Huber, J. Chem. Phys. **15**, 275 (1947).
[i] O. Steil, Zeits. f. physik. Chemie **B31**, 343 (1936).
[j] See reference 11.

[10] H. O. Kneser, Ann. d. Physik. **39**, 261 (1941); J. Acous. Soc. Am. **5**, 122 (1933).
[11] W. T. Richards and J. A. Reid, J. Chem. Phys. **2**, 193 (1934); W. T. Richards **4**, 561 (1936).
[12] H. O. Kneser and V. O. Knudsen, Ann. der Phys. **21**, 682 (1935).
[13] A. Eucken and S. Aybar, Zeits. f. physik. Chemie **B46**, 195 (1940); A. Eucken and R. Becker, *ibid.* **27**, 219 (1934).

Numerical Determination of Fundamental Modes

DONALD A. FLANDERS AND GEORGE SHORTLEY*
Argonne National Laboratory, Chicago, Illinois
(Received July 12, 1950)

A convenient and practical method of numerical determination of the fundamental eigenfunction and eigenvalue in a class of linear eigenvalue problems has been developed and applied in two and three dimensions. The method is based on use of a network and of difference equations, but departs from previous methods in that it is not iterative. Rather, a polynomial operator is applied to a trial function just once, to accomplish a determinable degree of reduction in all eigenfunctions other than the fundamental that are contained in the trial function. In the case of the diffusion equation, the polynomial operator is a Tschebyscheff polynomial of a simple averaging operator. It is shown that this operator, when of degree m, is "better" than any other polynomial operator of this degree and much "better" than m iterations of a simple averaging operator—"better" in the sense of accomplishing to a greater degree the elimination of *all* unwanted eigenfunctions. Techniques for the use of computing equipment for application of the polynomial operator are discussed. By orthogonalization, the method can be applied to modes other than the fundamental.

INTRODUCTION

IN connection with work on complex problems occurring in certain applications of diffusion theory, a convenient and practical method of numerical determination of the fundamental eigenfunction and eigenvalue in linear eigenvalue problems has been developed and applied in two and three dimensions.[1]

While the methods we have developed are applicable to a wide class of systems of linear partial differential equations and a wide variety of linear boundary conditions, we feel that the new ideas and techniques will be most useful if they are first presented in the simplest possible setting. Hence, in this paper we shall first consider in detail the numerical determination of the lowest eigenvalue and corresponding fundamental eigenfunction of the equation

$$\Delta u+\alpha u=0, \tag{1}$$

which occurs in the theories of heat conduction, diffusion, vibration, and elsewhere. For simplicity we shall fix attention on a connected two-dimensional region bounded by straight-line segments parallel to the x- and y-coordinate axes, and shall consider the simplest boundary condition

$$u=0 \qquad \text{on the boundary.} \tag{2}$$

The methods discussed can be generalized so as to apply to more general linear differential equations of elliptic type, to systems of such equations, to more complex geometrical configurations, to more general linear boundary and interface conditions and to more dimensions. Systems of equations will be discussed briefly in the last section.

We replace the continuum by a square lattice of mesh h, and equation (1) by a difference equation, in the usual way. Again for simplicity we assume that the dimensions are such that the boundaries fall along rows of lattice points, and that a lattice point falls at the origin of coordinates. We further assume that the mesh is so small that every pair of interior points can be joined by a polygonal path consisting of segments of the network lying wholly in the interior of the region; we shall say that the *interior* points are *connected* by the net. We denote the lattice points by P_{ij} and the corresponding values of u by u_{ij}, where $i=x/h, j=y/h$.

If we use the notation ωu_{ij} to represent the average of the values of u at the four nearest neighbors of the point P_{ij}, i.e.,

$$\omega u_{ij}=\tfrac{1}{4}(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}),$$

the second-difference approximation[2] to Δu at the point P_{ij} is

$$\Delta u_{ij}\approx\frac{\omega u_{ij}-u_{ij}}{\tfrac{1}{4}h^2}.$$

The difference equation that approximates (1) is then

$$\omega u_{ij}-u_{ij}+\tfrac{1}{4}h^2\alpha u_{ij}=0,$$

or

$$\omega u_{ij}=\lambda u_{ij}, \tag{3}$$

where

$$\lambda=1-\tfrac{1}{4}h^2\alpha. \tag{4}$$

* Operations Research Office, The Johns Hopkins University, on leave of absence from The Ohio State University; consultant at the Argonne National Laboratory.

[1] For a review of previous work on numerical methods in two and three dimensions (mostly work on boundary-value rather than eigenvalue problems) see Thomas J. Higgins, *Numerical Methods of Analysis in Engineering* (The Macmillan Company, New York, 1949), Chapter 10. This chapter contains an extensive bibliography.

[2] Nothing we say in this paper will be changed except in minor details if one chooses to use the higher-order approximation to Δu that is obtained if we define the averaging operator ω as

$$\omega u_{ij}=\tfrac{1}{5}(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1})+\tfrac{1}{20}(u_{i+1,j+1}+u_{i-1,j+1}+u_{i+1,j-1}+u_{i-1,j-1}),$$

in which case

$$\Delta u_{ij}\approx(\omega u_{ij}-u_{ij})/\tfrac{3}{10}h^2$$

and (4) becomes

$$\lambda=1-\tfrac{3}{10}h^2\alpha.$$

This expression for Δu_{ij} is valid for any seventh-order polynomial passing through the nine points, whereas the expression in the text is valid for a cubic polynomial passing through the five points. For a detailed discussion of approximations to the Laplacian, see Shortley and Weller, J. App. Phys. 9, 345 (1938).

The highest eigenvalue λ_1 of (3) will give an approximation to the lowest eigenvalue α_1 of (1) according to the relation (4), and the corresponding eigenvector of (3) will approximate to the fundamental eigenfunction of (1). We shall not discuss in this paper the accuracy of these approximations; it is well known that the solutions of the difference equation converge to solutions of the differential equation as $h \to 0$.

We shall consider the function u_{ij} to be defined only at interior points of the region. At a point adjacent to the boundary ωu_{ij} then will be taken to represent $\frac{1}{4}$ the sum of the values at the three, two, or one neighboring interior points, since one or more zero boundary values occur in the averaging process. We assume that there are N interior points, and for convenience number them from 1 to N, replacing the double index i,j by a single index k. The set of values u_k will then be denoted by a column vector $\mathbf{u}=(u_k)$. For the sake of brevity two components of $\mathbf{u}$ will be called neighboring if they represent values of u at neighboring points of the lattice.

Let $\omega_{kl}=\frac{1}{4}$ or 0 according as u_k and u_l are neighbors or not, and let ω denote the square matrix with elements ω_{kl}. The matrix equation

$$\omega\mathbf{u}=\lambda\mathbf{u} \tag{5}$$

then represents the result of incorporating the boundary conditions into the system of Eqs. (3). Thus Eq. (5) is the complete finite-difference analog of the analytic system (1), (2). It is the algebraic eigenvalue problem set by (5) with which we shall be concerned henceforth.

BASIC PROPERTIES OF THE SOLUTIONS OF THE MATRIX EQUATION

The general theory of equations such as (5) is well known. We set down here such general facts as we shall need, and we derive certain further useful properties specific to the particular equation.

The matrix ω is real and symmetric. Hence, it has N real eigenvalues and a complete orthonormal system of eigenvectors.[3] We denote the eigenvalues by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N,$$

and the corresponding eigenvectors of an arbitrary (but fixed) complete orthonormal system by

$$\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^N. \tag{6}$$

(Since $-\mathbf{u}$ is an eigenvector belonging to the eigenvalue λ if $\mathbf{u}$ is, and is normalized if $\mathbf{u}$ is, we shall assume throughout that whenever $\mathbf{u}$ denotes an eigenvector, the maximum of the absolute values of the components of $\mathbf{u}$ is equal to the value of some component u_k of $\mathbf{u}$.)

The averaging operation performed by ω cannot increase the maximum of the absolute values of the components of any eigenvector $\mathbf{u}$, so that every $|\lambda_n| \leq 1$. In fact one can readily see that only the inequality can hold because of the zero boundary condition. (It is clear that the $\leq$ sign will hold for more general operators than ω provided the corresponding matrix contains no negative elements and the sum of the elements in every row is ≤ 1, while the inequality may be asserted if in addition the sum of the elements of some row is <1.)

The highest eigenvalue, λ_1, is non-degenerate (i.e., $\lambda_1 > \lambda_2$ or, equivalently, there are not two linearly independent eigenvectors with eigenvalue λ_1), *and* $\mathbf{u}^1$ *is everywhere positive.* These facts may be proved by using the variational analog of the eigenvalue problem, which may be stated thus: If $\mathbf{u}^0$ is a normalized eigenvector of (5) with eigenvalue λ_0, then $\mathbf{u}^0$ is a vector that gives the quadratic form

$$(\mathbf{u}, \omega\mathbf{u}) \equiv \sum_{k,l} \omega_{kl} u_k u_l$$

the stationary value λ_0, subject to the normalization condition

$$(\mathbf{u}, \mathbf{u}) \equiv \sum_k u_k^2 = 1;$$

and conversely.

Let $\mathbf{v}$ be any normalized eigenvector belonging to λ_1 (i.e., with eigenvalue λ_1), and let $\mathbf{w}=(|v_k|)$. Then $\mathbf{w}$ is normalized, and we shall show that it is also an eigenvector belonging to λ_1. For since no $\omega_{kl}<0$ it follows that $\omega_{kl}w_k w_l \geq \omega_{kl}v_k v_l$ in every case, and hence that $(\mathbf{w}, \omega\mathbf{w}) \geq \lambda_1$. Since λ_1 is the maximum value that $(\mathbf{u}, \omega\mathbf{u})$ can assume for any normalized vector $\mathbf{u}$, we

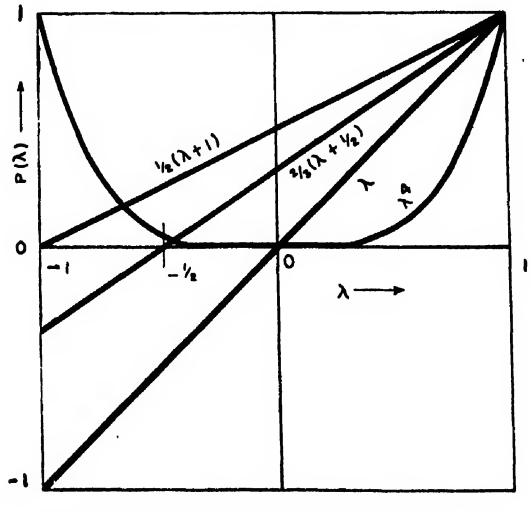


FIG. 1. The factors $P(\lambda)$ corresponding to certain polynomial operators $P(\omega)$.

[3] If ω is a real but unsymmetric matrix, and weight factors γ_i exist such that $\gamma_i\omega_{ij}=\gamma_j\omega_{ji}$, ω will have N real eigenvalues and a complete system of eigenvectors orthonormal with respect to the weight factors γ_i, the direct product being defined by

$$(\mathbf{u}, \mathbf{v}) = \Sigma_i \gamma_i u_i v_i.$$

In this case $(\mathbf{u}, \omega\mathbf{u})$ is a symmetric quadratic form that is an extremum for an eigenvector. This case is of importance when different choices of net spacing are made in different regions, the weight factor representing essentially the area of continuum associated with each net point. The weight factors for change of net spacing are given by Kimball and Shortley, Phys. Rev. 45, 815 (1934). The techniques of the present paper are readily applicable to matrix operators of this type, and in particular to the Schrödinger problem discussed by Kimball and Shortley.

TABLE I. Reduction factor $1/T_m(d)$, where $d=(2/r)-1$.

	$r=0.9$ $d=1.22222$	0.925 1.16216	0.95 1.10526	0.96 1.08333	0.97 1.06186
$m=$ 1	0.81818	0.86047	0.90476	0.92308	0.94175
2	0.50311	0.58781	0.69290	0.74227	0.79676
3	0.27499	0.35816	0.47961	0.54477	0.62361
4	0.14490	0.20884	0.31588	0.38023	0.46502
6	0.03393	0.06890	0.12996	0.17425	0.24138
8	0.01061	0.02229	0.05251	0.07792	0.12123
12	0.00077	0.00238	0.00852	0.01542	0.03001
16	0.00006	0.00025	0.00138	0.00305	0.00740
24			0.00004	0.00012	0.00045
32					0.00003

	$r=0.98$ $d=1.04082$	0.985 1.03046	0.99 1.02020	0.9925 1.01511	0.995 1.01005
$m=$ 1	0.96078	0.97044	0.98020	0.98511	0.99005
2	0.85719	0.88993	0.92454	0.94259	0.96117
3	0.72066	0.77799	0.84264	0.87814	0.91603
4	0.58076	0.65559	0.74637	0.79932	0.85846
6	0.35076	0.43397	0.55043	0.62750	0.72281
8	0.20285	0.27373	0.38606	0.46941	0.58347
12	0.06555	0.10396	0.17853	0.24514	0.35360
16	0.02101	0.03892	0.08052	0.12382	0.20514
24	0.00215	0.00543	0.01620	0.03098	0.06686
32	0.00022	0.00076	0.00325	0.00772	0.02149
40	0.00002	0.00011	0.00065	0.00193	0.00692
48		0.00001	0.00013	0.00048	0.00223
56			0.00003	0.00012	0.00072
64			0.00001	0.00003	0.00023
72					0.00008
80					0.00002

must have $(\mathbf{w}, \omega\mathbf{w})=\lambda_1$, and hence $\mathbf{w}$ is an eigenvector belonging to λ_1. Thus $\omega w_k \equiv \sum_l \omega_{kl} w_l = \lambda_1 w_k$ for every k.

Now $\mathbf{v}$ (and hence $\mathbf{w}$) can have no zero component. For if $\mathbf{w}$ has any zero values, there must be at least one point P_r for which $w_r=0$ that has a neighboring point P_s for which $w_s>0$. This follows because $\mathbf{w}$ does not vanish identically and because the interior points are connected by the net. But we have arrived at a contradiction since on the one hand $\omega w_r=\lambda_1 w_r=0$, while on the other ωw_r is a sum of non-negative terms $\omega_{rl} w_l$ of which at least one, $\omega_{rs} w_s$, is not zero.

Further, no component of $\mathbf{v}$ is negative. As before one could find neighboring values $v_r<0, v_s>0$ ($v_k=0$ having been excluded). But now we have $(\mathbf{w}, \omega\mathbf{w})>(\mathbf{v}, \omega\mathbf{v})$ since $\omega_{rs} w_r w_s = -\omega_{rs} v_r v_s > 0 > \omega_{rs} v_r v_s$.

Finally, no two linearly independent eigenvectors can belong to λ_1. For if $\mathbf{v}^1$ and $\mathbf{v}^2$ were such eigenvectors, normalized and with only positive components, their difference would be an (unnormalized) eigenvector belonging to λ_1 that would have both positive and negative values, which is impossible. Thus $\mathbf{u}^1$ is unique and is everywhere positive. Since every eigenvector belonging to a different eigenvalue must be orthogonal to $\mathbf{u}^1$, $\mathbf{u}^1$ must be the only everywhere positive eigenvector in the set $\mathbf{u}^1, \cdots, \mathbf{u}^N$.

We have thus proved that

$$1>\lambda_1>\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_N>-1. \tag{7}$$

An arbitrary vector $\mathbf{v}$ has a unique expansion in terms of the set (6):

$$\mathbf{v}=\sum c_n \mathbf{u}^n, \tag{8}$$

where c_n is the inner product of $\mathbf{v}$ and $\mathbf{u}^n$, that is

$$c_n=(\mathbf{v}, \mathbf{u}^n)=\sum_k v_k u^n{}_k.$$

If $\mathbf{v}$ is everywhere positive the coefficient c_1 must also be positive, and if furthermore $\mathbf{v}$ is reasonably pillow-shaped, c_1 will probably be larger than the magnitude of any other coefficient.

The effect of applying the operator ω to an arbitrary vector is to produce the vector

$$\omega\mathbf{v}=\sum_n \lambda_n c_n \mathbf{u}^n.$$

More generally, if $P(\omega)$ is any polynomial in ω,

$$P(\omega)\mathbf{v}=\sum_n P(\lambda_n) c_n \mathbf{u}^n,$$

that is to say that each coefficient c_n is multiplied by the value of the polynomial $P(\lambda)$ at $\lambda=\lambda_n$. Figure 1 shows the graphs of the factors $P(\lambda)$ corresponding to the operators ω, ω^4 and $(\omega-a)/(1-a)$ for $a=-1$ and for $a=-\frac{1}{2}$.

It is clear from Fig. 1 that the operator ω will reduce the magnitude of every coefficient c_n in proportion to the nearness of the corresponding eigenvalue to 0, but that if $|\lambda_n|$ is close to λ_1 the reduction relative to c_1 will be small, or there may even be a relative increase if $|\lambda_n|>\lambda_1$. Iteration of the operator ω, represented by the operator ω^p, is seen to have the same properties, except that the relative effect increases with increasing p. On the other hand the operator $(\omega-a)/(1-a)$ can be chosen with a negative so that every $c_n(\neq c_1)$ is reduced in magnitude, not only absolutely, but relatively to the magnitude of c_1.

TECHNIQUES FOR THE SOLUTION OF THE MATRIX EQUATION

By iterative application to a trial vector (8) of a polynomial operator $P(\omega)$ of the type shown in Fig. 1 it is possible to obtain a sequence of vectors $\mathbf{v}^0(=\mathbf{v})$, $\mathbf{v}^1, \cdots, \mathbf{v}^i, \cdots$, such that the ratios $|c_n{}^i/c_1{}^i|$ are successively reduced. Past procedures for obtaining the fundamental solution of such a matrix equation have been of this type. p iterations of a polynomial operator $P_d(\omega)$ of degree d are equivalent to the operator $[P_d(\omega)]^p$, which is of degree pd. We shall now show that for given degree m there exists a "best" polynomial operator $P_m(\omega)$. Since this polynomial is not a power of any polynomial of lower degree, it is therefore "better" than the result of p iterations of any operator of degree d such that $pd\leq m$.

The criterion for the "best" polynomial operator $P(\omega)$ of degree m may be rigorously expressed in the following form: We assume that the eigenvalues of the

problem lie in an interval $(-a, c)$ and that there is an interval $(-a, b)$, with $b<c$, within which it is desired that the maximum of $|P(\lambda)/P(c)|$ shall be a minimum. By limiting consideration to those polynomials for which $P(c)=1$ we then seek among such polynomials those for which the maximum of $|P(\lambda)|$ is a minimum throughout the interval $(-a, b)$. If such a polynomial exists for each choice of m, a, b, c, then by taking $a=-\lambda_N$, $b=\lambda_2$, $c=\lambda_1$ we can preserve the magnitude of c_1 in the expansion of an arbitrary trial vector (8), and obtain the maximum reduction in magnitude of the remaining coefficients.[4] In practice, we do not know in advance what values should be assigned to a, b, and c. In later sections we shall show how such assignments may reasonably be made.

The existence of the desired "best" polynomial is given by the following theorem, in which, for convenience, we have made the changes of variable

$$\Omega=\frac{2}{a+b}\omega+\frac{a-b}{a+b}, \qquad \mu(\lambda)=\frac{2}{a+b}\lambda+\frac{a-b}{a+b} \tag{9}$$

[so that $\mu(-a)=-1$, $\mu(b)=+1$], and have set $d=\mu(c)\equiv 2(a+c)/(a+b)-1>1$.

THEOREM: Among all polynomials of degree m in μ having the value $+1$ at $\mu=d>1$, there is just one having the minimum maximum absolute value throughout the interval $(-1, +1)$, namely the polynomial

$$S_m(\mu)=T_m(\mu)/T_m(d), \tag{10}$$

where $T_m(\mu)$ is the mth-order Tschebyscheff polynomial,[5] obtained by expanding

$$T_m(\mu)=\cos(m \arccos\mu)$$

in powers of μ.

PROOF: $T_m(\mu)$ is alternately $+1$ and -1 at the $m+1$ points $\mu_k=\cos(k\pi/m)$ $(k=0, 1, \cdots, m)$. Consequently $|S_m(\mu)|$ has $1/T_m(d)$ as its maximum at these $m+1$ values of μ. Suppose that $R(\mu)$ is a polynomial of degree m that is $=1$ at $\mu=d$ and has equal or smaller maximum absolute value in the interval $(-1, +1)$. Let $Q(\mu)=S_m(\mu)-R(\mu)$. Then $Q(\mu)$ is of degree m or less, $Q(d)=0$, and $Q(\mu_k)\geq 0$ or ≤ 0 according as k is even or odd. We proceed to show that $Q(\mu)$ has at least $m+1$ zeros in the closed interval $[-1, d]$, which is possible only if $Q\equiv 0$, i.e., if $R\equiv S_m$. Note that if $0<k<m$ and $Q(\mu_k)=0$ then μ_k is a multiple zero since both S_m and R must have natural extrema at μ_k, and hence $S_m'(\mu_k)-R'(\mu_k)=Q'(\mu_k)=0$. Now consider the interval $[\mu_k, \mu_{k+1}]$ $(k=1, \cdots, m-2)$. Either Q is positive at one end of this interval and negative at the other, in which case there is a zero in the interior of the interval; or Q is zero at one or both ends of the interval. In the latter case Q has at least two zeros in the closed interval—let us assign *one* of these zeros to the interval $[\mu_k, \mu_{k+1}]$ and one to the adjoining interval that has the double zero at its end point. In this way we see that Q has at least one zero assignable to each of the $m-2$ intervals mentioned above. The intervals $[\mu_0, \mu_1]$ and $[\mu_{m-1}, \mu_m]$ have (at least) either an interior zero, a double zero at μ_1 or μ_{m-1}, or a single zero at μ_0 or μ_m. Hence at least one zero is assignable to each of the m intervals $[\mu_k, \mu_{k+1}]$ $(k=0, \cdots, m-1)$. Q also has a zero at d. Hence it has at least $m+1$ zeros, which proves that $Q\equiv 0$ and $R\equiv S_m$.

The quantity $1/T_m(d)$ is the maximum of $|S_m(\mu)|$ in the interval $[-1, +1]$, and is therefore the measure of the reduction of the coefficients of all eigenfunctions with eigenvalue in the range $-a<\lambda<b$, relative to an eigenfunction with eigenvalue c, when $S_m(\Omega)$ is applied to a trial vector (8). Table I gives this reduction factor, and can be used to determine the degree of the polynomial required to remove unwanted eigenfunctions to any given degree of accuracy. The parameter d has been replaced by the quantity $r=(a+b)/(a+c)$, which is more directly related to the spacing of the eigenvalues.

Figure 2 illustrates the effect of the operator $S_6(\Omega)$ for the case $a=1$, $b=0.9$, $c=1$ $(r=0.95)$. All eigenfunctions with eigenvalues in the range from -1 to $+0.9$ are multiplied by a factor of 0.13 or less, as shown by the curve $S_6[(20\lambda+1)/19]$. For comparison the sixth order curve λ^6 is shown, and the range over which this is less than 0.13 is indicated. In an actual computation

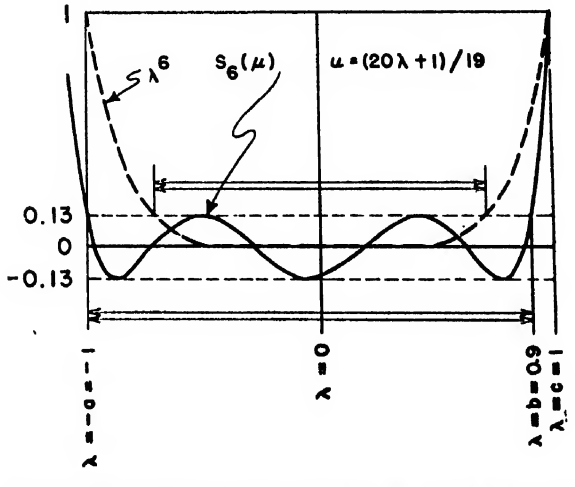


FIG. 2. As shown by the double arrows, the Tschebyscheff operator $S_6(\Omega)$ will accomplish a given reduction over a much wider range than will the operator ω^6. (In the upper right-hand corner, u should be μ.)

[4] In one sense the "best" polynomial would be $(\omega-\lambda_2)(\omega-\lambda_3)\cdots(\omega-\lambda_N)$, since this would eliminate every eigenvector except $\mathbf{u}^1$. But in any practical case such as we envisage, where the number of eigenvectors is large (in the hundreds or thousands), this would not actually be as useful (even assuming the eigenvalues known) as the polynomial we describe, since it would be of much higher degree.

[5] We are indebted to Drs. Tukey and Grosch for suggesting the possible usefulness of Tschebyscheff polynomials, in the course of a discussion of this work at the IBM Seminar on Scientific Computation held in Endicott, New York in November, 1949. We had previously used polynomials with zeros equally spaced. These were fairly efficient, but appreciably less so than the Tschebyscheff polynomials.

a polynomial of order higher than six would always be used.

In the practical application of an operator $S_m(\Omega)$ to a trial vector it is often convenient to use the following factor formula for Tschebyscheff polynomials:

$$T_{pq}(x)=\prod_{k=1}^{p}\left[T_q(x)-\cos\frac{2k-1}{p}\frac{\pi}{2}\right]. \tag{11}$$

This follows from the identity

$$\cos(pq\arccos x)\equiv\cos\{p\arccos[\cos(q\arccos x)]\},$$

which may be written

$$T_{pq}(x)=T_p[T_q(x)].$$

Since $T_p(y)$ has p zeros located at $y=\cos[(2k-1)/p](\pi/2)$ $(k=1,\cdots,p)$, (11) results from setting $y=T_q(x)$ in the factored form of $T_p(y)$.

If we have a vector v that is known to be a linear combination of just the first two eigenvectors, i.e.,

$$\mathbf{v}=c_1\mathbf{u}^1+c_2\mathbf{u}^2, \tag{12a}$$

one can determine λ_1, λ_2, $\mathbf{u}^1$, $\mathbf{u}^2$ by the following "harmonic analysis": First compute $\omega\mathbf{v}$ and $\omega^2\mathbf{v}$,

$$\omega\mathbf{v}=\lambda_1c_1\mathbf{u}^1+\lambda_2c_2\mathbf{u}^2, \tag{12b}$$

$$\omega^2\mathbf{v}=\lambda_1^2c_1\mathbf{u}^1+\lambda_2^2c_2\mathbf{u}^2. \tag{12c}$$

Next form the inner products

$$\left.\begin{aligned}(\mathbf{v},\mathbf{v})\equiv A&=c_1^2+c_2^2,\\(\mathbf{v},\omega\mathbf{v})\equiv B&=\lambda_1c_1^2+\lambda_2c_2^2,\\(\omega\mathbf{v},\omega\mathbf{v})\equiv(\mathbf{v},\omega^2\mathbf{v})\equiv C&=\lambda_1^2c_1^2+\lambda_2^2c_2^2,\\(\omega\mathbf{v},\omega^2\mathbf{v})\equiv D&=\lambda_1^3c_1^2+\lambda_2^3c_2^2.\end{aligned}\right\} \tag{13}$$

When c_1^2 and c_2^2 are eliminated from the equations (13) there result two independent equations for λ_1 and λ_2, which may be written in the form

$$\begin{vmatrix}A&1&1\\B&\lambda_1&\lambda_2\\C&\lambda_1^2&\lambda_2^2\end{vmatrix}=\begin{vmatrix}B&\lambda_1&\lambda_2\\C&\lambda_1^2&\lambda_2^2\\D&\lambda_1^3&\lambda_2^3\end{vmatrix}=0. \tag{14}$$

If these be solved for λ_1 and λ_2, and the possibilities $\lambda_1=0$, $\lambda_2=0$, $\lambda_1=\lambda_2$ (which do not concern us here) be discarded, the two λ's will be found to be the roots of the equation

$$\begin{vmatrix}A&B&1\\B&C&\lambda\\C&D&\lambda^2\end{vmatrix}=0. \tag{15}$$

$c_1\mathbf{u}^1$ and $c_2\mathbf{u}^2$ can then be determined from (12a, b) by

$$\left.\begin{aligned}c_1\mathbf{u}^1&=\frac{1}{\lambda_1-\lambda_2}\omega\mathbf{v}-\frac{\lambda_2}{\lambda_1-\lambda_2}\mathbf{v},\\c_2\mathbf{u}^2&=\frac{-1}{\lambda_1-\lambda_2}\omega\mathbf{v}+\frac{\lambda_1}{\lambda_1-\lambda_2}\mathbf{v}.\end{aligned}\right\} \tag{16}$$

If v is only approximately represented by the linear combination (12a), so that relations (13) are only approximately true, the solutions of (15) will give approximate values, λ_1' and λ_2', of λ_1 and λ_2. Since λ_1' and λ_2' $(\lambda_1'>\lambda_2')$ satisfy (13) exactly, $\lambda_1'A\equiv\lambda_1'(c_1^2+c_2^2)>\lambda_1'c_1^2+\lambda_2'c_2^2\equiv B$, $\lambda_1'>B/A\equiv(\mathbf{v},\omega\mathbf{v})/(\mathbf{v},\mathbf{v})$, i.e., λ_1' is greater than the estimate of λ_1 that would be obtained directly from the variational principle. In the examples we have tried we have always found $\lambda_1'<\lambda_1$, but we have not been able to prove that this will be true in general.

APPLICATION OF THE TECHNIQUES

In this section we shall discuss in some detail the application of the techniques developed in the preceding section. The general idea is first to do enough preliminary work to determine the rough shape of the eigenvector and approximate values of λ_1 and λ_2; then to make reasonable choices of a, b, and c, compute $r=(a+b)/(a+c)$, and choose an order m from Table I that will give adequate elimination of unwanted eigenfunctions. Then the operator $S_m(\Omega)$ given by (9) and (10) is applied to the rough eigenfunction to accomplish this elimination.

The problems for which these techniques were developed were based on relatively complicated regions that necessitated the use of hundreds or thousands of points to get even a moderately acceptable representation of the region and its boundary by the network. This has meant that the computations were possible only with the assistance of fairly high-speed computing machinery with a large memory, such as IBM punched-card equipment, particularly the 604 electronic computer. The following remarks are assumed to be concerned with such large-scale problems computed with comparable equipment.

In starting it is generally worth while to use a rather coarse net, even at the expense of considerably oversimplifying the geometry; e.g., if the region contains holes on whose boundaries the function is 0, it may be worth while to select a mesh so large that such a hole is represented by a single point at which the function is 0. The particular advantages of such a rough first approximation are that (a) the eigenvalues are much less densely spaced, so that approximate values of λ_1 and λ_2 are much more easily determined; (b) relation (4) shows that if λ' and λ'' are the eigenparameters for networks of mesh h' and h'', respectively, then λ_1'' and λ_2'' may be approximated from λ_1' and λ_2' by using $\lambda''=1-(h''/h')^2(1-\lambda')$; (c) the simplest possible interpolation in the vector derived from the coarse net will generally yield a vector that differs from the fundamental for the fine net principally by eigenvectors with eigenvalues near -1 (i.e., high-frequency oscillations), which are easy to eliminate. Thus, one may get a good starting vector for the fine network by treating the solution for the coarse network as a stepfunction in the fine network and smoothing it by a few preliminary operations with $(\omega+1)/2$, about equal in number to the ratio of h' to h''. A better starting func-

tion would be obtained by using an interpolation based on the difference equation (3) itself. Thus if 1, 2, 3, 4 are four points at the corners of a square of diagonal $2h'$, the difference equation determines the value u_0 at the center of the square as

$$u_0=(1/\lambda_1')\,\tfrac{1}{4}(u_1+u_2+u_3+u_4),$$

where λ_1' is the eigenvalue appropriate to a net of mesh h'. If we start with a net of mesh h, we can first use $h'=h/\sqrt{2}$ to fill in the center points of the meshes, then $h'=h/2$ to fill in the remaining points in a net of mesh $\frac{1}{2}h$.

The form of the starting function for the coarse network is not too important, provided the generally convex character of the function is observed. Advantage should be taken of any symmetries in the figure to reduce the number of points, which not only reduces the number of eigenfunctions but also decreases the density of the eigenvalues. It should be noted that the matrix resulting from using only a portion of a symmetrical figure is no longer symmetrical. For example, if the full figure is symmetrical along a diagonal and half the figure is used, the neighbors of a diagonal point are the neighbors on one side counted twice. The eigenvectors of the resulting matrix will not then be orthogonal in general. Their orthogonality will be restored if they are treated as eigenvectors of the matrix associated with the complete figure, which is done simply by giving to each point in the subregion actually used a weight equal to the number of points in the full region for which it stands.

In problems of the type considered here one may assume $a=1$ throughout, since $\lambda_N\approx-1$; and $c=1$ initially, since $\lambda_1\approx+1$. The first step is to pick an operator Ω and a polynomial $S_m(\Omega)$ of low degree that will enable one to compute λ_1 and λ_2 by the crude harmonic analysis of the previous section with fair accuracy. Thus, with $a=c=1$ the choice of $r=0.9$ and $m=8$ corresponds to $b=0.8$, $d=1.222$, $\Omega=(10\omega+1)/9$, $1/T_m(d)=0.0106$. This $S_m(\Omega)$ will reduce the coefficients of all eigenvectors in the expansion of v with eigenvalues lying in the range $-1\leq\lambda\leq0.8$ by at least the factor 0.01, and those with eigenvalues in the range $0.8\leq\lambda\leq1$ at least in proportion to the ratio $(\lambda-0.8)/0.2$. Ordinarily this will permit a harmonic analysis that will determine λ_1 and λ_2 with sufficient accuracy to choose a final operator for this network. However, if the number of points in the coarse network is rather large, it may be worth while to take r and m larger.

In the actual performance of an operation $S_m(\Omega)$ one is faced with a choice of methods. On the one hand $S_m(\Omega)$v may be computed by iterating with the operator Ω and then forming the polynomial as a linear combination of the results. This has the practical disadvantage that with large m the sizes of the terms are so disparate that a great deal of accuracy is lost in taking differences of large multiples of the results of the iterations. At the other extreme, $S_m(\Omega)$ may be factored into its linear factors,

$$S_m(\Omega)=\prod_{k=1}^{m}\frac{\Omega-\mu_k}{d-\mu_k},\qquad\left(\mu_k=\cos\frac{2k-1}{m}\frac{\pi}{2}\right)\tag{17}$$

and these linear operators applied successively. This method suffers from the disadvantage that the coefficients of the operator must be changed at each step, which is a dangerous source of error in machine operation. A reasonable practical compromise combines the two methods by use of the factor formula (11). Thus if we set

$$S_m(\Omega)=\prod_{k=1}^{p}\frac{T_q(\Omega)-\cos[(2k-1)/p]\pi/2}{T_q(d)-\cos[(2k-1)/p]\pi/2},\tag{18}$$

where q is about 6 or 8[6] and $p=m/q$, we may iterate q times with Ω, follow this with the computation of the first of the p polynomials, iterate on the result q times with Ω, compute the second polynomial factor, and so on. Since the cosine term in the denominator is small compared to $T_q(d)$, it may be omitted. This has the effect of changing the scale of the vector slightly, which is of no theoretical importance. If this omission is made, then the operator used in the iterations is always the same, while the coefficients used in the successive polynomials differ only in the constant term. Loss of significant figures is avoided by keeping q relatively small. When q is at least as large as 6 or 8, the change of scale produced by neglecting the cosine term in the denominator is not bothersome. (If the function does grow out of bounds, cutting it in half once will generally be sufficient. Note that this should be done *after* a polynomial factor is computed, not in the course of a set of iterations preparatory to the computation of such a polynomial.)

When a satisfactory solution for the coarse network has been obtained, a starting vector for the fine network may be constructed along the lines suggested above. At this point procedure will depend to some extent on whether the interest in the problem centers in the determination of the fundamental eigenvalue or of the fundamental eigenfunction. If the former, then b may safely be chosen somewhat less than λ_2 and we may determine λ_1 by harmonic analysis. If λ_1 and λ_2 are widely separated this method may also be used for determining the vector $c_1\mathbf{u}^1$. However, the determination of $c_1\mathbf{u}^1$ from (16) generally involves great loss of accuracy because of the nearness of λ_1 and λ_2, so if interest is centered in the eigenfunction, one is forced to choose b close to λ_2 with consequent increase in the size of m.

HIGHER MODES

While the method we have outlined is peculiarly adapted to obtaining the fundamental mode, it is pos-

[6] T_6 and T_8, which we have found most useful, are given by

$$2^{-5}T_6(\Omega)=\Omega^6-\tfrac{3}{2}\Omega^4+\tfrac{9}{16}\Omega^2-\tfrac{1}{32}$$
$$2^{-7}T_8(\Omega)=\Omega^8-2\Omega^6+\tfrac{3}{4}\Omega^4-\tfrac{1}{4}\Omega^2+\tfrac{1}{128}.$$

sibly useful for obtaining a limited number of the higher modes. Once the fundamental has been accurately obtained for the network, a trial vector for the next mode can be chosen that is accurately orthogonal to the fundamental. Application of $P(\omega)$ to such an orthogonal vector should not reintroduce (except as a result of rounding-off errors) any of the fundamental. Hence it should be possible to obtain a higher mode by orthogonalizing the trial vector to the predetermined lower modes, applying a suitable Tschebyscheff operator, and re-orthogonalizing as necessary. One useful application of this technique might be in the solution of Schrödinger's equation for diatomic molecules.

MORE GENERAL PROBLEMS

Several of the problems we have solved by a procedure similar to the above have been of a more general type than the one proposed in the introduction. Since the modifications introduced by this generalization may assist others in applying the procedure, we shall state briefly the most important features of the more general problems and their solution.

The single dependent variable u is replaced by two functions u and v that are to satisfy simultaneous equations

$$\left.\begin{aligned}(\Delta-E)u+Fv&=-\alpha u\\ Gu+(\Delta-H)v&=-\alpha v.\end{aligned}\right\} \tag{19}$$

The fundamental region is divided into subregions within each of which the coefficients E, F, G, H are constant functions of a parameter ρ. Certain homogeneous conditions are imposed across the interfaces. It is desired to determine ρ so that the fundamental eigenvalue $\alpha_1=0$, and to find the corresponding pair of eigenfunctions u, v.

The matrix equation analogous to (5) is then of the form

$$\begin{pmatrix}\omega-\frac{1}{4}h^2E & \frac{1}{4}h^2F\\ \frac{1}{4}h^2G & \omega-\frac{1}{4}h^2H\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}=\lambda\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}, \tag{20}$$

where $\lambda=1-\frac{1}{4}h^2\alpha$, and consequently it is desired that $\lambda_1=1$. As in the first problem the introduction of the boundary conditions (and here of the interface conditions as well) results in modification of the interpretation of the operator ω at certain points, but ω still represents a linear combination of neighboring values with non-negative coefficients whose sum is ≤ 1. The two vectors $\mathbf{u}$, $\mathbf{v}$ are considered to form a single vector for the purposes of the computation.

Note that the matrix is no longer symmetric. This means that the eigenvectors may not be assumed to be orthogonal. Hence it is necessary ultimately to compute the adjoint solution in order to form correct inner products if a harmonic analysis is to be carried out. In the early stages this is not necessary since it can be shown that the λ_1 and λ_2 determined by the analog of (15) represent the result of fitting relations (12) by least squares. The λ's thus determined are, however, less accurate than those determined by true inner products.

In these problems h is chosen sufficiently small so that the coefficients h^2E, h^2F, h^2G, h^2H are small compared to 1. Hence the operator is nearly the operator $\begin{pmatrix}\omega 0\\ 0\omega\end{pmatrix}$. Aside from the asymmetry the principal effect of the more general matrix is to increase the range of the eigenvalues. Since it is possible to estimate ρ (on physical grounds) with fair accuracy in advance, λ_1 is still approximately $+1$. λ_N, however, may be considerably less than -1, say -3 or -4 if $\mathbf{u}$ and $\mathbf{v}$ are disparate in size. Our first problem is then to estimate λ_N. This may be done by choosing a sequence of operators of the type $(\omega-a)/(1-a)$ to replace ω in the matrix. If one takes $a=0$ and applies the resulting operator two or three times, the presence of eigenvectors with eigenvalue <-1 will be revealed by their coefficients being multiplied by negative factors of large absolute value, which thus produces oscillations of increasing amplitude in the successive vectors. (That such eigenvectors will appear, provided $\lambda_N<-1$, is practically guaranteed by the rounding-off errors introduced into the computation.) When the existence of such large negative eigenvalues has been established, similar operations may be carried out with $a=-1$, $-\frac{3}{2}$, -2, etc., until the phenomenon disappears. Thus a practical value of a for use in determining the operator Ω in the Tschebyscheff polynomial is achieved. The polynomial operator S_m is now not a function of Ω but of the matrix operator in (20) with Ω substituted for ω.

A final word about the treatment of the interface conditions. The particular conditions imposed are expressed in the difference equations by assertions of the form: the value of u at an interface point is a weighted average of certain neighboring values of u in the adjacent regions. It would be possible to eliminate all interface points from the vector in a manner similar to that in which the boundary points were eliminated. However, since λ_1 is very nearly $+1$ in these problems it is more convenient to retain the interface points, replacing their values at each stage by the weighted average specified in the interface conditions. This averaging then becomes part of the operator. Since we are doing the averaging and determining the interface values always one step late, this procedure is only exact for an eigenfunction of eigenvalue unity.

We are indebted to many of the members of the Theoretical Physics Division of the Argonne Laboratory, and in particular to Drs. Frank Hoyt and Elmer Eisner, for helpful discussions during the development of these techniques.

Determination of Eigenvalues Using a Generalized Laplace Transform*

MORRIS D. FRIEDMAN
N.A.C.A., Ames Aeronautical Laboratory, Moffett Field, California
(Received August 14, 1950)

The classical eigenvalue problems of mathematical physics are solved by means of a Laplace Transform extended, not over the interval $(0, \infty)$ but over the interval of interest for the differential equation. The method is applied to the Hermite, Laguerre and Bessel equations and to the equation for the hypergeometric polynomials which include the Legendre, Tschebyscheff, and Jacobi polynomials as special cases.

INTRODUCTION

IN many fields of mathematical physics, for example, quantum mechanics, electro-magnetics, acoustics, there occur second-order differential equations depending on a parameter λ which have satisfactory or well-behaved solutions only for particular values of λ. A solution is said to be well behaved if it is single-valued, continuous, and finite throughout the configuration space of the system (i.e., for all values of x which x may assume). In case the range of the variable x is infinite an extra condition about the behavior of the solution at infinity is needed. The values of the parameter λ for which the solutions are well behaved functions are called eigenvalues and the corresponding solutions eigenfunctions. This paper proposes a general method for finding the eigenvalues of a second-order differential equation.

The usual method of solving these second-order equations, called the Sommerfeld Polynomial Method, is to substitute a power series expansion (whose form is determined by an asymptotic solution of the differential equation) for the unknown function and then to determine the coefficients by equating like powers of x. This gives the general solution of the equation. Then the radius of convergence of the power series is investigated in order to find for which values of λ the solution is well behaved. This method is long, and it will be shown that by using a generalized Laplace transform,[1] the same results can be achieved in a more direct way and with less labor.

Those equations which can be put into the Sturm-Liouville[2] form will be treated. The Bessel, Hermite, Laguerre, Legendre, Jacobi, Tschebyscheff, and the Trigonometric equations will be special cases. These equations can be put into the following general form:

$$[P(x)U'(x)]'-[Q(x)-\lambda R(x)]U(x)=0. \quad (1)$$

Since the solutions of (1) have singularities at the zeros of $P(x)$ or at ∞, the problem is to find values of λ for which $U(x)$ is well behaved at these singular points. The method proposed is as follows.

Multiply each term of Eq. (1) by $e^{-sx}dx$ and integrate over a suitably chosen contour. In general, the contour will be a line joining the singularities of the differential equation, that is, either the real axis between two zeroes of $P(x)$, or the part of the real axis connecting a zero of $P(x)$ to ∞. The fact that $U(x)$ is well behaved at the singular points implies that the transformed function must be a regular, analytic function of s. Those values of λ for which the transform is analytic are the eigenvalues.

Consider

$$\int_C [P(x)U'(x)]'e^{-sx}dx - \int_C [Q(x)-\lambda R(x)]U(x)e^{-sx}dx=0. \quad (2)$$

When the first term on the left is integrated by parts twice, there results

$$\{P(x)U'(x)e^{-sx}+sP(x)U(x)e^{-sx}\}_C + s^2\int_C P(x)U(x)e^{-sx}dx - \int_C [sP'(x)+Q(x)-\lambda R(x)]U(x)e^{-sx}dx=0. \quad (3)$$

Now, if that contour is selected which will make the braces vanish, there results the integral equation:

$$\int_C [s^2P(x)-sP'(x)-Q(x)+\lambda R(x)]U(x)e^{-sx}dx=0, \quad (4)$$

or dividing by $R(x)$

$$\int_C \left[s^2\frac{P(x)}{R(x)}-s\frac{P'(x)}{R(x)}+\frac{Q(x)}{R(x)}+\lambda\right]U(x)R(x)e^{-sx}dx=0. \quad (5)$$

Suppose now that $P(x)/R(x)$, $P'(x)/R(x)$ and $Q(x)/R(x)$ are polynomials. This is actually the case in the standard eigenvalue problems. Put

$$U(x)R(x)=\bar{U}(x), \quad (6)$$

---

* Submitted in partial fulfillment of the requirements for the Master of Science degree.

[1] L. Pauling and E. B. Wilson, *Introduction to Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1935), first edition.

[2] R. Courant and D. Hilbert, *Methoden der Mathematische Physik* (J. Springer, Berlin—Photo-lithoprint, Interscience Publications, New York, 1937), Vol. I.

and let

$$\int_C \bar{U}(x)e^{-sx}dx = U(s), \tag{7}$$

where $U(s)$ is a generalized Laplace transform of $\bar{U}(x)$. Under certain conditions, multiplication of the integral in (7) by a power of x is equivalent to multiplying the derivative to that order by (-1) to that power; e.g.

$$\int x\bar{U}(x)e^{-sx}dx = -v'(s),$$

$$\int x^2\bar{U}(x)e^{-sx}dx = v''(s), \text{ etc.}^3$$

Therefore, Eq. (6) may be represented as the result of a function of s and the differential operator with respect to s acting on the integral of $\bar{U}(x)e^{-sx}dx$, or

$$g(s, D)\int_C \bar{U}(x)e^{-sx}dx = g(s, D)v(s) = 0, \tag{8}$$

where D is the differential operator with respect to s and $v(s)$ is the Laplace transform of $U(x)$. The result is, thus, a differential equation in $v(s)$ depending upon the parameter λ. The method consists in finding those values of λ for which $v(s)$ is an analytic function of s.

EIGENVALUES OF THE HYPERGEOMETRIC EQUATION

The method will first be applied to the particular case of the hypergeometric equation whose form is[4]

$$[(1+x)^r(1-x)^{p+q-r+1}U'(x)]' - \lambda(1+x)^{r-1}(1-x)^{p+q-r}U(x) = 0. \tag{9}$$

From this equation can be derived the Jacobi ($r=P+1$; $q=0$; $p=Q+P+1$), the Tschebyscheff ($r=\frac{1}{2}$; $p=q=0$), and the Legendre ($r=1$, $p+q=r$) equations. It will be shown that substitution of these values in the solution of λ for the hypergeometric equation will yield the corresponding solutions for the eigenvalues of the respective equations.

Since $P(x)$ in (9) has zeroes at $x=+1$ and $x=-1$, the terms of (9) are multiplied by e^{-sx} and the result is integrated between the limits of -1 and $+1$:

$$\int_{-1}^{1}[(1+x)^r(1-x)^{p+q-r+1}U'(x)]'e^{-sx}dx - \int_{-1}^{1}\lambda(1+x)^{r-1}(1-x)^{p+q-r}U(x)e^{-sx}dx = 0, \tag{10}$$

which can be written as

$$\int_{-1}^{1}[s^2(1+x)^r(1-x)^{p+q-r+1} - sr(1+x)^{r-1}(1-x)^{p+q-r+1} + s(p+q-r+1)(1+x)^r(1-x)^{p+q-r} + \lambda(1+x)^{r-1}(1-x)^{p+q-r}][U(x)e^{-sx}dx] = 0, \tag{11}$$

the same form as (5). Now, let

$$\int_{-1}^{1}(1+x)^{r-1}(1-x)^{p+q-r}U(x)e^{-sx}dx = v(s). \tag{12}$$

Recalling that multiplication of the transform by a power of x is equivalent to differentiation, (11) then becomes

$$s^2v - s^2v'' - srv - srv' + s(p+q-r+1)v - s(p+q-r+1)v' + \lambda v = 0, \tag{13}$$

where $v=v(s)$ and v' and v'' are derivatives of $v(s)$ with respect to x.

From (12) it is seen that if $U(x)$ is well behaved then $v(s)$ is an analytic function of s for all values of s. It is now necessary to find for which values of λ, (13) has analytic solutions. If $v(s)$ is analytic, then it can be written as

$$v(s) = \sum_0^\infty a_k s^k; \text{ and similarly,}$$

$$v'(s) = \sum_0^\infty k a_k s^{k-1}; \text{ and } v''(s) = \sum_0^\infty k(k-1)a_k s^{k-2}.$$

When these series are substituted in (13), there results

$$\sum a_{k-2}s^k - \sum (k^2-k)a_k s^k - r\sum a_{k-1}s^k + (p+q-r+1)\sum a_{k-1}s^k - (p+q-r+1)\sum k a_k s^k + \sum \lambda a_k s^k = 0, \tag{14}$$

or

$$(k^2+kp+kq-\lambda)a_k - a_{k-2} - (p+q-r+1)a_{k-1} = 0; \text{ for } k=0, 1, 2, \cdots. \tag{15}$$

(*Note:* $a_{-1}=a_{-2}=0$.)

Suppose, now, that a_n is the first non-zero coefficient. Then for $k=n$, (15) becomes

$$(n^2+np+nq-\lambda)a_n = 0, \tag{16}$$

so that λ must be

$$\lambda = n^2+np+nq, \tag{17}$$

and then the n-th eigenvalue is

$$\lambda_n = n(n+p+q). \quad n=1, 2, \cdots.$$

Notice that if a_n is the first non-zero coefficient then $v(0)=v'(0)=\cdots=v^{(n-1)}(0)=0$; or

$$(-1)^k v^k(0) = \int_{-1}^{1}(1+x)^{r-1}(1-x)^{p+q-r}U(x)x^k dx = 0 \qquad (k=0, 1, \cdots, n-1), \tag{18}$$

[3] R. V. Churchill, *Modern Operational Methods in Engineering* (McGraw-Hill Book Company, Inc., New York, 1944), first edition.

[4] E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge University Press, New York, 1947). (Note: In this paper, all equations were put in the self-adjoint form before solving. Differentiating the self-adjoint form and dividing by the common factor will bring the equation to the form listed in the reference.)

so that $U(x)$, the solution corresponding to the n-th eigenvalue is orthogonal to x^k $(k=0, 1, \cdots, n-1)$ with the weight function $(1+x)^{r-1}(1-x)^{p+q-r}$. This contains the well-known orthogonality property of the Jacobi, Tschebyscheff, and Legendre functions.[2,5]

If those values of p and q which were mentioned before are substituted in (17), the eigenvalues for the Jacobi, Tschebyscheff, and Legendre equations are a result.

The Legendre equation of higher order[2]

$$[(1-x^2)U'(x)]'-h^2(1-x^2)^{-1}U(x)+\lambda U(x)=0, \quad (19)$$

can be put in a more manageable form by the substitution of $Z(x)=(1-x^2)^{\frac{1}{2}h}U(x)$. The resulting equation

$$[(1-x^2)^{h+1}Z'(x)]'-[h(h+1)-\lambda](1-x^2)^hZ(x)=0, \quad (20)$$

transforms into the differential equation

$$s^2v''(s)+2s(h+1)v'(s)-(s^2+h(h+1)+\lambda)v(s)=0. \quad (21)$$

From (21), it is determined that

$$\lambda=n^2+n+2nh+h^2+h, \quad (h=0, 1, 2, \cdots).$$

THE CONFLUENT HYPERGEOMETRIC EQUATION

Another class of equations, the confluent hypergeometric, is allied to the hypergeometric mentioned above. The equations of this new class which will be treated are the Hermite, Laguerre, and Bessel.[2,4]

The Hermite equation, in self-adjoint form, is[2]

$$[e^{-x^2}U'(x)]'+\lambda e^{-x^2}U(x)=0. \quad (22)$$

Solutions are desired such that $e^{-x^2}U(x)$ is zero at $x=\pm\infty$.

$$\int_{-\infty}^{\infty}[e^{-x^2}U'(x)]'e^{-sx}dx+\int_{-\infty}^{\infty}\lambda e^{-x^2}U(x)e^{-sx}dx=0. \quad (23)$$

If the first integral on the left is integrated by parts; twice:

$$\left\{e^{-x^2-sx}U'(x)+se^{-x^2-sx}U(x)\right\}_{-\infty}^{\infty}$$
$$+s\int_{-\infty}^{\infty}(s+2x)e^{-x^2-sx}U(x)dx$$
$$+\int_{-\infty}^{\infty}\lambda e^{-x^2-sx}U(x)dx=0. \quad (24)$$

Now, around the chosen contour, if $U(x)$ is well behaved, the braces are zero and

$$v(s)=\int_{-\infty}^{\infty}e^{-x^2-sx}U(x)dx,$$

is substituted in (25), recalling that the product of a power of x and the integrand in the integral for which $v(s)$ is being substituted is a derivative of $v(s)$, then

$$s^2v(s)-2sv'(s)+\lambda v(s)=0. \quad (25)$$

This equation can be solved by separation of variables

$$v'(s)/v(s)=(s^2+\lambda)/2s-s/2+\lambda/2s, \quad (26)$$

or

$$\log v=s^2/4+(\lambda\log s)/2. \quad (27)$$

And, therefore,

$$v(s)=s^{\frac{1}{2}\lambda}\exp(\tfrac{1}{4}s^2). \quad (28)$$

It is concluded that in order for $v(s)$, determined in (27), to be analytic λ must be $2n$, where n is a positive integer.

It will now be shown that the well behaved solutions of the Hermite equation are polynomials. Assume that $\lambda=2n$ and that the corresponding solution of $U(x)$ may be written as

$$U(x)=e^{x^2}[e^{-x^2}V(x)]',$$

then (25) becomes

$$V(s)=\int_{-\infty}^{\infty}e^{-sx}[e^{-x^2}V(x)]'dx. \quad (29)$$

After an integration by parts,

$$V'(s)=e^{-sx-x^2}V(x) \quad +s\int_{-\infty}^{\infty}e^{-sx-x^2}V(x)dx, \quad (30)$$

so that from (28)

$$\int_{-\infty}^{\infty}e^{-sx-x^2}V(x)dx=s^{n-1}\exp(s^2/4). \quad (31)$$

By induction, it is seen that if

$$U(x)=e^{x^2}\frac{d^n}{dx^n}[e^{-x^2}W(x)]; \quad (32)$$

then

$$\int_{-\infty}^{\infty}e^{-sx-x^2}W(x)dx=\exp(s^2/4). \quad (33)$$

But

$$\int_{-\infty}^{\infty}e^{-sx-x^2}dx$$
$$=\int_{-\infty}^{\infty}\exp[-(x+\tfrac{1}{2}s)^2]\exp(s^2/4)dx$$
$$=\int_{-\infty}^{\infty}\exp(s^2/4)\exp(-y^2)dy=\pi^{\frac{1}{2}}\exp(\tfrac{1}{4}s^2), \quad (34)$$

so that

$$\int_{-\infty}^{\infty}e^{-sx-x^2}(dx/\pi^{\frac{1}{2}})=\exp(\tfrac{1}{4}s^2), \quad (35)$$

[5] P. Frank and R. Von-Mises, *Differentiale Gleichungen—Vol. I* (Mary S. Rosenberg, New York, 1943), second edition.

and $W(x)$ may be taken equal to $\pi^{-\frac{1}{2}}$ to satisfy Eq. (33). This shows that

$$U(x)=\frac{e^{x^2}}{\pi^{\frac{1}{2}}}\frac{d^n}{dx^n}(e^{-x^2}). \tag{36}$$

The confluent hypergeometric equation also contains, as a special case, the Laguerre equation.[2,4]

$$[xe^{-x}U'(x)]'+\lambda e^{-x}U(x)=0. \tag{37}$$

A well behaved solution of this equation is one which is regular at $x=0$ and such that $e^{-x}U(x)$ vanishes at $x=\infty$. Just as for the Legendre equation, so too, for the Laguerre there is an equation for functions of higher order. This equation

$$[x^{k+1}e^{-x}U'(x)]'+\lambda x^k e^{-x}U(x)=0, \tag{38}$$

can be solved in the same manner as the equation of first order.

Multiply (38) by e^{-sx} and integrate along C, the real axis from 0 to ∞. Then

$$\int_C [xe^{-x}U'(x)]'e^{-sx}dx+\int_C \lambda e^{-x}e^{-sx}U(x)dx=0. \tag{39}$$

This Eq. (39) is similar in form to Eq. (2) and since the contour selected will cancel all integrated terms, (40) can immediately be put in the form (5) as follows:

$$s^2\int_C xe^{-x-sx}U(x)dx+s\int_C xe^{-x-sx}U(x)dx$$

$$-s\int_C e^{-x-sx}U(x)dx+\int_C \lambda e^{-x-sx}U(x)dx=0. \tag{40}$$

Now let the transform be

$$v(s)=\int_0^\infty e^{-x(1+s)}U(x)dx, \tag{41}$$

then (41) becomes

$$-s^2v'(s)-sv'(s)-sv(s)+\lambda v(s)=0. \tag{42}$$

After collection of terms and separation of variables, there results

$$\frac{v'(s)}{v(s)}=\frac{-s}{s(s+1)}=\frac{\lambda}{s}-\frac{1+\lambda}{s+1}, \tag{43}$$

and

$$\log v=\lambda\log s-(1+\lambda)\log(s+1), \tag{44}$$

or

$$v(s)=s^\lambda/(s+1)^{1+\lambda}. \tag{45}$$

Since $U(x)$ behaves like $e^{x-\epsilon}$ at $x=\infty$, $v(s)$ must be regular at $x=0$. Were λ negative, $v(s)$ would have a pole; were λ a fraction there would be a branch point, and since both assumptions are contrary to the conditions, then $\lambda=n$, a positive integer.

(Note that the transform of x^n is $n!/(1+s)^{n+1}$; so if $v(s)$ be expanded in inverse powers of $(1+s)$, $U(x)$ can be found.)

From (45), for $\lambda=n$,

$$v(s)=(1+s-1)^n/(1+s)^{n+1}=\frac{1}{s+1}\left(1-\frac{1}{s+1}\right)^n, \tag{46}$$

and

$$v(s)=\sum_0^n \frac{(-1)^k\binom{n}{k}}{(1+s)^{k+1}}. \tag{47}$$

The inverse transform of $v(s)$ in (47) is[2]

$$U(x)=\sum_0^n (-1)^k\frac{x^k}{k!}\frac{n!}{k!(n-k)!}, \tag{48}$$

which is a solution of (37).

The Bessel equation[6]

$$x^2U''(x)+(2x+2)U'(x)-\lambda U(x)=0,$$

can also be put into the self-adjoint form with the weight function $\exp(-2/x)$. The result is

$$[\exp(-2/x)x^2U'(x)]'-\lambda\exp(-2/x)U(x)=0, \tag{49}$$

and the more general form

$$[\exp(-b/x)x^aU'(x)]'-\lambda\exp(-b/x)x^{a-2}U(x)=0, \tag{50}$$

which are also special cases of the confluent hypergeometric equation.

A solution single-valued on the unit circle $|x|=1$ is desired. Again, multiply the equation, in this case (49), by e^{-sx} and integrate around C, the unit circle, obtaining

$$\int_C [\exp(-2/x)x^2U'(x)]'e^{-sx}dx$$

$$-\int_C \lambda\exp(-2/x)e^{-sx}U(x)dx=0. \tag{51}$$

Integrating as before, yields the result in the form of (4)

$$\{\exp(-2/x)x^2e^{-sx}U'(x)+s\exp(-2/x)x^2e^{-sx}U(x)\}_C$$

$$+\int_C (s^2x^2-2sx+2s-\lambda)$$

$$\times\exp(-2/x)e^{-sx}U(x)dx=0. \tag{52}$$

Now, around this contour, because $U(x)$ is single-valued, the braces evaluate to zero. Then, putting

$$v(s)=\int_C \exp(-s/x)e^{-sx}U(x)dx, \tag{53}$$

[6] J. Kroll and O. Frink, "A new class of orthogonal polynomials," Abstract 249, Bull. Am. Math. Soc., July 1948.

the resulting equation is

$$s^2v''(s)+2sv'(s)+2sv(s)-\lambda v(s)=0. \quad (54)$$

Substituting a power series for $v(s)$, and equating coefficients of like powers of s, yields

$$n^2-n+2n=\lambda, \quad (55)$$

or

$$\lambda=n^2+n.^5$$

In Eq. (50), λ can be determined to be $n(n-a+1)$.[6]

The property of orthogonality demonstrated for the Jacobi equation can, similarly, be shown to apply to the other equations treated.

Finally, consider the special case of the Sturm-Liouville type equation, the Trigonometric[2]

$$U''(x)+\lambda^2U(x)=0, \quad (56)$$

with the conditions $U(0)=U(1)=0$.

Multiply (56) by e^{-sx} and integrate from 0 to 1. Thus

$$\int_0^1 U''(x)e^{-sx}dx+\int_0^1 \lambda^2U(x)e^{-sx}dx=0. \quad (57)$$

Integrate the first integral by parts twice:

$$\left\{U'(x)e^{-sx}+sU(x)e^{-sx}\right\}_0^1+s^2\int_0^1 U(x)e^{-sx}dx$$

$$+\int_0^1 \lambda^2U(x)e^{-sx}dx=0. \quad (58)$$

Evaluate the brace and set

$$v(s)=\int_0^1 U(x)e^{-sx}dx, \quad (59)$$

obtaining

$$U'(1)e^{-s}-U'(0)+(s^2+\lambda^2)v(s)=0. \quad (60)$$

Solving for $v(s)$

$$v(s)=\frac{U'(0)-U'(1)e^{-s}}{s^2+\lambda^2}. \quad (61)$$

Now, for $v(s)$ to be analytic, it is necessary to get rid of the poles at $s=\pm i\lambda$. Therefore the numerator of (61) must be zero when $s=\pm i\lambda$.

Inserting $\pm i\lambda$ for s and solving simultaneously

$$\begin{aligned} U'(0)-U'(1)e^{-i\lambda}&=0,\\ U'(0)-U'(1)e^{+i\lambda}&=0, \end{aligned} \quad (62)$$

$$U'(1)(e^{i\lambda}-e^{-i\lambda})=0. \quad (63)$$

Since $U'(1)$ is not zero, the parenthesis must be, and therefore

$$\sin\lambda=0,$$

or

$$\lambda=n\pi; \quad \lambda^2=n^2\pi^2.$$

SUMMARY

A method which can be used to determine the eigenvalues of equations of mathematical physics has thus been demonstrated on the equations of the Sturm-Liouville type. The method also demonstrates that the solutions have certain orthogonality properties and in the special cases of the Hermite and Laguerre equations allows the solution to be found.

Letters to the Editor

Comment on "Relationship between Gough-Joule Coefficients and Moduli of Vulcanized Rubbers"[1]

M. Mooney
United States Rubber Company, Passaic, New Jersey
August 31, 1950

In the article of the title quoted above the authors were interested in establishing by experiment the theoretical equality between the tangent shearing modulus of a rubber and the ratio of the temperature coefficient of shearing stress to the negative temperature coefficient of shearing strain. Their published data can be used also for calculating the value of an important property, the relative temperature coefficient of relative stress, σ, at constant strain, γ, or $T/\sigma(\partial\sigma/\partial T)_\gamma$. The value of this coefficient is unity if there is entropy change but no energy change in an isothermal strain, as is postulated in the kinetic theory of elasticity.

The values of the coefficient can be calculated from the data given by Conant, Hall, and Thurman for σ/γ, $1/\gamma(\partial\sigma/\partial T)_\gamma$, and T, the reference temperature.

The calculations yield the results:

Compound	$T/\sigma(\partial\sigma/\partial T)_\gamma$
Hevea *B*	0.82
GR-S	1.33
Butaprene	0.90
Butyl	1.05
Neoprene	1.04
Hevea *A*	0.81
Hevea *C*	0.67

While the values of the coefficient are correct in order of magnitude, they depart from unity in many cases by amounts which are presumably larger than the experimental error. These results are qualitatively in agreement with those published by Copeland and Mooney,[2] who also found significant departures from the theoretical value, unity, in several cases. Quantitatively, however, the results of the two groups of experimentalists are not in good agreement.

Obviously more work in the field of thermo-elasticity is called for to determine whether the differences in results are caused by the use of different compounds, different test methods, or other factors. However, it is growing more and more evident that, contrary to the kinetic theory of elasticity, the energy of an elastomer subjected to isothermal strain sometimes contributes appreciably to the stress.

Concerning the experimental method used by Conant and associates, it may be remarked, incidentally, that the constraints on the thermal expansion imposed by a sandwich type of sample involves some corrections which may be appreciable, and certainly are not easily estimated.

[1] F. S. Conant, G. L. Hall, and G. R. Thurman, J. App. Phys. 20, 526 (1949).
[2] L. E. Copeland and M. Mooney, J. App. Phys. 19, 450 (1948).

An Approximation Method for the Determination of the Elastic Constants of Cubic Single Crystals

J. R. Neighbours and Charles S. Smith
Case Institute of Technology, Cleveland, Ohio
Received October 5, 1950

A recent investigator[1] has shown how the elastic constants of cubic crystals may be found from measurement of the three wave velocities in an arbitrary direction in the crystal. It is the purpose of this paper to show how this approximation method may be extended to higher orders.

Materials having cubic symmetry are characterized by three elastic constants, C_{11}, C_{12}, C_{44}. These constants relate stress and strain in a particularly simple way when the reference axes, XYZ, are also the crystallographic axes. In another set of axes, $X'Y'Z'$, the relationship is more complicated.

In general, the relationship between the sets of axes is governed by the scheme of direction cosines.

	X	Y	Z
X'	l_1	m_1	n_1
Y'	l_2	m_2	n_2
Z'	l_3	m_3	n_3

(1)

If the Y' axis is arbitarily taken to lie in the $X-Y$ plane, this scheme may be written in terms of l_1, m_1, and n_1 only.

	X	Y	Z
X'	l_1	m_1	n_1
Y'	$-m_1[l_1^2+m_1^2]^{-\frac{1}{2}}$	$l_1[l_1^2+m_1^2]^{-\frac{1}{2}}$	0
Z'	$-l_1n_1[l_1^2+m_1^2]^{-\frac{1}{2}}$	$-m_1n_1[l_1^2+m_1^2]^{-\frac{1}{2}}$	$[l_1^2+m_1^2]^{\frac{1}{2}}$

(2)

It may be shown that, for irrotational plane waves traveling in the X' direction, the equations of motion are:

$$\left.\begin{aligned}\rho\frac{\partial^2u'}{\partial t^2}&=C_{11}'\frac{\partial^2u'}{\partial x'^2}+C_{16}'\frac{\partial^2v'}{\partial x'^2}+C_{15}'\frac{\partial^2w'}{\partial x'^2},\\ \rho\frac{\partial^2v'}{\partial t^2}&=C_{16}'\frac{\partial^2u'}{\partial x'^2}+C_{66}'\frac{\partial^2v'}{\partial x'^2}+C_{56}'\frac{\partial^2w'}{\partial x'^2},\\ \rho\frac{\partial^2w'}{\partial t^2}&=C_{15}'\frac{\partial^2u'}{\partial x'^2}+C_{56}'\frac{\partial^2v'}{\partial x'^2}+C_{55}'\frac{\partial^2w'}{\partial x'^2},\end{aligned}\right\} \tag{3}$$

where

$$\left.\begin{aligned}C_{11}'&=C_{11}-2[l_1^2m_1^2+l_1^2n_1^2+m_1^2n_1^2]C,\\ C_{16}'&=l_1m_1[m_1^2-l_1^2][l_1^2+m_1^2]^{-1}C,\\ C_{66}'&=C_{44}+2l_1^2m_1^2[l_1^2+m_1^2]^{-1}C,\\ C_{15}'&=-n_1[l_1^4+m_1^4-n_1^2(l_1^2+m_1^2)][l_1^2+m_1^2]^{-1}C,\\ C_{55}'&=C_{44}+2n_1^2[l_1^4+l_1^2m_1^2+m_1^4][l_1^2+m_1^2]^{-1}C,\\ C_{56}'&=l_1m_1n_1[l_1^2-m_1^2][l_1^2+m_1^2]^{-1}C,\\ C&=[C_{11}-C_{12}-2C_{44}],\end{aligned}\right\} \tag{4}$$

and u', v', w' are the displacements in the X', Y', Z' directions, respectively.

The secular equation obtained from Eq. (3) is

$$\begin{vmatrix}C_{11}'-\rho V^2 & C_{16}' & C_{15}'\\ C_{16}' & C_{66}'-\rho V^2 & C_{56}'\\ C_{15}' & C_{56}' & C_{55}'-\rho V^2\end{vmatrix}=0. \tag{5}$$

The roots of this equation may be found by the perturbation method to be

$$\left.\begin{aligned}\rho V_1^2&=C_{11}'+\left\{\left[\frac{C_{15}'^2}{C_{11}'-C_{55}'}+\frac{C_{16}'^2}{C_{11}'-C_{66}'}\right]+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{11}'-C_{66}')(C_{11}'-C_{55}')}\right]+\cdots\right\},\\ \rho V_2^2&=C_{66}'+\left\{\left[\frac{C_{16}'^2}{C_{66}'-C_{11}'}+\frac{C_{56}'^2}{C_{66}'-C_{55}'}\right]+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{66}'-C_{11}')(C_{66}'-C_{55}')}\right]+\cdots\right\},\\ \rho V_3^2&=C_{55}'+\left\{\left[\frac{C_{15}'^2}{C_{55}'-C_{11}'}+\frac{C_{56}'^2}{C_{55}'-C_{66}'}\right]+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{55}'-C_{11}')(C_{55}'-C_{66}')}\right]+\cdots\right\}.\end{aligned}\right\} \tag{6}$$

Combining the perturbation roots of the secular determinant with the first three of Eqs. (4) results in the set of equations:

$$\left.\begin{aligned}
C_{11}' &= C_{11}-2[l_1^2m_1^2+l_1^2n_1^2+m_1^2n_1^2]C \\
&= \rho V_1^2-\left\{\left[\frac{C_{15}'^2}{C_{11}'-C_{55}'}+\frac{C_{16}'^2}{C_{11}'-C_{66}'}\right]\right. \\
&\quad \left.+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{11}'-C_{55}')(C_{11}'-C_{66}')}\right]+\cdots\right\}, \\
C_{66}' &= C_{44}+2l_1^2m_1^2[l_1^2+m_1^2]^{-1}C \\
&= \rho V_2^2-\left\{\left[\frac{C_{16}'^2}{C_{66}'-C_{11}'}+\frac{C_{56}'^2}{C_{66}'-C_{55}'}\right]\right. \\
&\quad \left.+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{66}'-C_{11}')(C_{66}'-C_{55}')}\right]+\cdots\right\}, \\
C_{55}' &= C_{44}+2n_1^2[l_1^4+l_1^2m_1^2+m_1^4][l_1^2+m_1^2]^{-1}C \\
&= \rho V_3^2-\left\{\left[\frac{C_{15}'^2}{C_{55}'-C_{11}'}+\frac{C_{56}'^2}{C_{55}'-C_{66}'}\right]\right. \\
&\quad \left.+\left[\frac{2C_{15}'C_{16}'C_{56}'}{(C_{55}'-C_{11}')(C_{55}'-C_{66}')}\right]+\cdots\right\}.
\end{aligned}\right\} \quad (7)$$

If the three wave velocities V_1, V_2, V_3 in the X' direction are determined, then Eqs. (7) may be solved by successive approximations to yield C_{11}, C_{12}, and C_{44}. The X' direction, specified by l_1, m_1, n_1, which also must be measured, need not be a special direction in the cubic crystal. Certain exceptional directions are noted below.

The values of the perturbation terms (within braces, Eq. 6) depend in complex ways upon the elastic constants and direction in the crystal, but roughly the absolute value of the quantity $C=[C_{11}-C_{12}-2C_{44}]$ determines the magnitude of these terms. For ionic crystals such as silver chloride where the quantity $|C|$ is small, the higher order perturbation terms are quite small.[1] Much larger values for corresponding perturbation terms are found for metallic crystals such as nickel where the quantity $|C|$ is about an order of magnitude larger. In our experience, about five trials are required in making a solution of (7) for a metallic crystal.

In general, a given set of direction cosines defines any one of six crystallographically equivalent directions in the first octant, and care must be exercised to choose an advantageous order for the direction cosines. Since n_2 has been chosen equal to zero, the perturbation approximations to the roots of the secular equation will be better approximations to the actual roots for certain choices of l_1, m_1, and n_1. In practice, the better choices are those that place the X' axis near an $[h\,h\,l]$ type direction.

It should be pointed out that the directions [100] and [111] are ambiguous in that for each of these directions the two transverse wave velocities are equal. All three elastic constants, C_{11}, C_{12}, and C_{44}, cannot be determined by wave velocity measurements in these directions. Wave velocity measurements in those general directions near to these two are suitable in principle but not in practice for a determination of all three elastic constants.

The [110] direction in the cubic crystal is, on the other hand, unique in that wave velocity measurements in this direction will determine all three elastic constants, and in this direction the perturbation terms are zero. General directions near to this special direction will yield good accuracy for the constants and only small perturbation corrections will be necessary.

The elastic constants numerically computed from (7) may be checked by using the results to make a numerical solution of the secular determinant, Eq. (5). The three values for the wave velocities thus obtained from (5) should agree with the three measured wave velocities.

It is felt that the method outlined here has several advantages over that of measuring the wave velocities in crystals cut in special, simple directions. In practice, many difficulties may be experienced in attempts to grow single crystals large enough to be cut to a desired orientation. Even if crystals of adequate size are grown, use of the above perturbation method may be quicker and simpler than cutting an oriented crystal. This perturbation method has been used by the authors for determining the elastic constants of nickel, which work will be published at an early date. This work has been supported by the ONR.

[1] D. L. Arenberg, J. App. Phys. 21, 941 (1950).

Coefficient of Self-Diffusion of Copper

G. Cohen and G. C. Kuczynski
Metallurgy Department, Polytechnic Institute of Brooklyn, Brooklyn, New York
August 31, 1950

It has been shown previously[1-3] that the coefficients of self-diffusion of metals can be determined by measuring, as a function of time and temperature, the width x of the "neck" formed during sintering of a metallic wire to a cylinder of the same metal. The method was described in detail elsewhere.[3]

The specimens were prepared from OFHC copper. The annealed copper wire of 0.005-inch diameter was wound around the rods of ¼-inch diameter and heated in dry hydrogen at temperatures and for time periods indicated in Table I. After heating, the specimens were cut longitudinally, polished, etched, and measured under the microscope. The coefficients of self-diffusion D were calculated from the equation[3]

$$D=3/80\pi\ kT/\delta^3\gamma t\ x^6/a^2,$$

where k is the Boltzmann constant, a the radius of the wire, T absolute temperature, t the time of heating, δ the interatomic

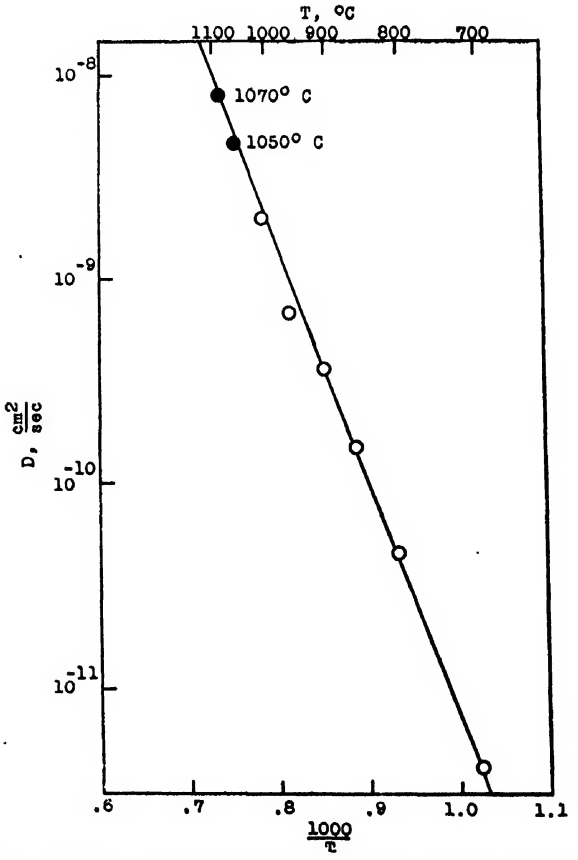


FIG. 1. Variation of self-diffusion coefficient of copper with temperature. Natural logarithm of D *versus* reciprocal of absolute temperature.

TABLE I. Self-diffusion coefficient D of copper.

Temperature, °C	Time (hrs)	$\times 10^{-4}$ cm	D cm²/sec.
700	4.0	6.1	4.06×10^{-12}
800	2.0	8.5	4.60×10^{-11}
850	1.0	9.4	1.53×10^{-10}
900	1.0	11.1	3.58×10^{-10}
950	1.0	12.2	6.80×10^{-10}
1000	0.5	13.2	1.95×10^{-9}

distance (in the present case 2.55A), and γ the surface energy of solid copper taken as 1370 ergs/cm².[4] The D-values so obtained are listed in Table I along with measured values of x. From the $\ln D/1000/T$ plot in Fig. 1, the coefficient of self-diffusion of copper was found to be expressed by the following equation: $D=4.1e^{-54,000/RT}$, which agrees fairly well with that obtained by J. Steigman, W. Shockley, and F. C. Nix,[5] which is $D=11.0e^{-57,200/RT}$. The points in Fig. 1 marked by black circles and corresponding to temperatures 1050°C and 1070°C were determined independently[6] by the same method. They fall very well on the straight line resulting from this investigation.

[1] G. C. Kuczynski, Trans. Am. Soc. Met. 55, 169 (1949).
[2] G. C. Kuczynski, Phys. Rev. 75, 1309 (1949).
[3] G. C. Kuczynski, J. App. Phys. 21, 632 (1950).
[4] H. Udin, A. Shaler, and J. Wulff, Trans. Am. Soc. Met. 55, 186 (1949).
[5] J. Steigman, W. Shockley, and F. C. Nix, Phys. Rev. 56, 13 (1939).
[6] B. Alexander, G. C. Kuczynski, and M. Dawson, "Relations between diffusion and viscous flow in metals," Symposium of the Physics of Powder Metallurgy, Sylvania Electric Products, Inc. (August 1949) (to be published soon).

Erratum: Transonic Potential Flow of a Compressible Fluid

[J. App. Phys. 21, 771–778 (1950)]

W. R. SEARS
Cornell University, Ithaca, New York

SINCE the publication of the above article, Professor Y. H. Kuo has changed his analysis of the stability of mixed potential flow. His conclusions are now different from those reported in my paper.

Specifically, Dr. Kuo discovered an inconsistency in the orders of magnitude of terms resulting from his boundary condition. Upon correction of this, the term k/m disappears from the definition of σ (see my Eq. (9)), and consequently the profound effect of the local slope of the body at the sonic line disappears from the results. His final conclusions, based on the revised calculation, are that accelerating transonic flow is stable with respect to the class of disturbances considered, while decelerating flow is unstable to compression disturbances, on convex surfaces regardless of the slope m. Obviously, this brings him into agreement with other authors, who have analyzed the problem from different standpoints, and seems to eliminate the interesting possibility of stable transonic potential flow sketched in my Fig. 10.

Erratum: Separation of Gases by Fractional Permeation through Membranes

[J. App. Phys. 21, 279–283 (1950)]

SOL WELLER AND WALDO A. STEINER
Synthetic Liquid Fuels Branch, Bureau of Mines, Pittsburgh, Pennsylvania

BY an error in transcription the sign of the quantity C has been given incorrectly in some of the equations in the above paper. C should be defined by the equation

$$C=\tfrac{1}{2}[1+(\alpha-1)p/P],$$

and Eqs. (12) and (13) should read

$$x=f(i)=Ai-C+[A^2i^2+(\alpha-2AC)i+C^2]^{\frac{1}{2}}, \tag{12}$$

$$\frac{dn_B}{n_B}=\frac{di}{f(i)-i}=\frac{di}{(A-1)i-C+[A^2i^2+2Bi+C^2]^{\frac{1}{2}}}, \tag{13}$$

where $2B=\alpha-2AC$. The subsequent equations and Fig. 2 are correct.

We are grateful to Dr. T. P. Wilson for pointing out this error to us.

Deformations and Stresses in Bourdon Tubes*

R. A. CLARK** AND E. REISSNER
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts

August 18, 1950

A BOURDON tube may be considered as a toroidal, elastic shell with thin walls, oval cross section, and with closed ends. When such a closed tube is exposed on the outside or on the inside to uniform normal wall pressure, the curvature of the center line of the tube changes proportionally with the applied wall pressure. Measurements of the resultant motion of one end of the tube with reference to the other may then be interpreted, after appropriate calibration, as pressure measurements.

The physical nature of the action of the tube may be understood by means of the following considerations. Take first a *complete* torus subject to uniform wall pressure. If the pressure is on the outside, it will produce tension in the part of the walls nearest to the axis of the torus and compression in the part of the walls furthest from the axis. This stress distribution then evidently produces an internal bending moment with axis parallel to the axis of the torus.

In the Bourdon tube *part* of the complete torus is taken and the ends are closed up. The shell is now no longer able to take the bending moment which occurred in the complete torus. Consequently one may consider the state of stress and deformation in the tube as a superposition of two states: (1) stresses and deformations in a complete torus subject to wall pressure; (2) stresses and deformations in an incomplete torus acted upon by bending moments at the ends.[1] The magnitude of the applied moment of state 2 is chosen such that it cancels the corresponding internal moment of state 1. It is the superimposed state 2 which is responsible for the change of curvature of the center line of the tube.

State 1 is treated as a problem in the theory of axi-symmetrical deformations of thin shells of revolution. State 2 is treated in the same manner after introduction into the theory of a non-axi-symmetrical component of circumferential displacement.[2]

If for the sake of discussion, the curved tube is considered as made up of two systems of curved rods, one system with circular center lines and the other system consisting of closed, oval rings, it is now understood that the action of the tube involves direct stresses and negligible bending of the circular rods and bending with negligible direct strains of the oval rings.[3–5]

There are two well-known limiting solutions both for state 1 and state 2. In the limiting solution for state 1 which concerns the complete torus under pressure the bending stresses in the oval rings are neglected and determination of the direct stresses becomes a problem of the membrane theory of thin shells.[6] In the limiting solution for state 2 the oval rings are considered rigid and the curved tube is stressed and deformed in accordance with the elementary theory of beams.

Combination of these two limiting states for a theory of the Bourdon tube has again been suggested quite recently.[7] It is, however, easy to see that joint application of the two limiting

solutions is not justified physically. The membrane solution for state 1 will be applicable only for tubes with sufficiently *thin* walls. The elementary-beam theory solution for state 2 will be applicable only for tubes with sufficiently *thick* walls. So far as is known, no results have yet been given concerning the meaning of the word "sufficiently" for the problem of state 1.

In order to gain further insight into the problem the authors have undertaken a study from the point of view of the theory of thin shells of the stresses and deformations of Bourdon tubes with elliptical cross section. In this theory the problem can be reduced to the determination of the periodic solutions of two simultaneous, linear, non-homogeneous second-order differential equations with periodic coefficients.

It is found that the solutions depend primarily on two dimensionless parameters which involve the dimensions of the tube. These parameters are of the form b/c and bc/ah, where a = radius of center line of the tube, h = wall thickness of the tube, c = major semi-axis of the elliptical cross section, and b = minor semi-axis of the elliptical cross section.

Quantitative solutions of the differential equations of the problem led to the result that the membrane solution for state 1 is adequate as soon as the parameter bc/ah is larger than about 10. On the other hand, the elementary beam-theory solution for state 2 ceases to be adequate as soon as bc/ah is larger than about 0.5.

Two different methods of solution were employed for the differential equations of the problem. For values of the parameter bc/ah up to about 5, trigonometric series were used. This led to infinite systems of simultaneous linear equations for the coefficients of the trigonometric series. These were solved approximately by retaining only a finite number of the coefficients. In this way up to ten simultaneous equations for ten unknowns were needed.

The other method of solution was an extension to *non-homogeneous* differential equations of what is often referred to as the W.K.B.J. method of asymptotic integration. This method is applicable for sufficiently large values of the parameter bc/ah. It was found that adequate results were obtained in this way as soon as bc/ah was larger than about 2. The consequent overlapping of the ranges of validity of the series solution and of the asymptotic solution permitted a check on the accuracy of either method.

Derivation of the asymptotic results proved to be very much more convenient than use of the series method. Remarkably simple explicit formulas could be obtained in the asymptotic range and these are listed below.

The sensitivity of the tube is given by the formula

$$\frac{\Delta a}{a}=\frac{p}{E}\frac{a}{b}\frac{c^2-b^2}{h^2}\left(1.63-\frac{ah}{cb}\right). \tag{1}$$

In Eq. (1) p is the applied wall pressure, E the modulus of elasticity of the tube material, and Δa the change of the radius a of the tube caused by the applied pressure. A small effect of Poisson's ratio is incorporated in the numerical coefficient 1.69 which is based on $\nu=\frac{1}{3}$. The general form of this coefficient is $[3(1-\nu^2)]^{\frac{1}{2}}$. Equation (1) is applicable as soon as cb/ah is larger than about 2. For very much larger values of cb/ah the second term inside the parenthesis becomes negligible and the result becomes still simpler.

The maximum stress in the tube is caused by bending of the elliptical rings and occurs at the ends of the major axis of the cross section. When cb/ah is larger than about 2, it is given by the formula

$$\sigma_m=1.88p(ah/cb)^{\frac{1}{2}}(c^2-b^2)/h^2. \tag{2}$$

It is further found that the stress σ_m is always smaller than the corresponding bending stress in a straight tube with the same elliptical cross section and with the same uniform wall pressure p. Denoting this stress for the straight tube by σ_{sm} it is found that, in the range of applicability of Eq. (2),

$$\frac{\sigma_m}{\sigma_{sm}}=\frac{1.87b_0}{b_0-b_1}\left(\frac{ah}{cb}\right)^{\frac{1}{2}}, \tag{3}$$

where b_0 and b_1 are expressed in terms of complete elliptic integrals in the form

$$\left.\begin{aligned} b_0&=(4/\pi)E(e), \quad e^2=1-(b/c)^2\\ b_1&=(8/3\pi e^2)[(1-e^2)K(e)-(1-\tfrac{1}{2}e^2)E(e)]\end{aligned}\right\}. \tag{4}$$

For values of bc/ah less than about 2, no simple explicit formula can be given for σ_m. An exception is the limiting relation

$$\lim_{bc/ah\to 0}\sigma_m/\sigma_{sm}=1. \tag{5}$$

Detailed derivations, together with numerical results, will be published in the near future.

* The following is based on work performed with the support of the ONR under Contract N5 ori-07834 with Massachusetts Institute of Technology.
** Now Department of Mathematics, Case Institute of Technology.

[1] In this formulation one has disregarded local effects at and near the ends of the tube which fall within the scope of St. Venant's principle.

[2] E. Reissner, *Bending of Curved Tubes*, Proc. Nat. Acad. Sci. U.S.A. **35**, 204–208 (1949).

[3] H. Lorenz, "Theorie der Röhrenfeder manometer," Physik. Zeits. **18**, 117 (1917).

[4] M. Tueda, "Mathematical theories of Bourdon pressure tubes and bending of curved pipes," Mem. Coll. Eng., Kyoto Imp. Univ. **8**, 102–115 (1934); **9**, 132–152 (1936).

[5] A. Wolf, "An elementary theory of the Bourdon gage," J. App. Mech. **13**, A207–A210 (1946).

[6] H. Lorenz, Theorie der Röhrenfeder manometer, Z. Verein Deutscher Ing. **54**, 1865 (1910).

[7] G. Schubert, *Über Effekte zweiter Ordnung bei Biegung und Torsion dünnwandiger Rohre elliptischen Querschmitts, Ing.*—Arch **12**, 53–63 (1941).

Shadowgrams of Spherical Missiles Entering Water at Supersonic Speeds

J. H. McMillen, R. L. Kramer, and D. E. Allmand
Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland
October 24, 1950

While a great deal is known about the flow of water when objects are propelled through it at modest speeds, very little is known about the flow or other physical responses of the water when objects are shot into it at supersonic speeds. In order to investigate the events which accompany water entry at supersonic speeds (above 4800 ft./sec.), a shadowgraphic investigation was carried out in which spherical missiles were shot into water at speeds of 7000 ft./sec. These speeds correspond to a Mach number of 1.46 in water. Shadowgraphic records have the advantage that they indicate the simultaneous (exposure times less than a microsecond) general disposition of the transient pressure field over the entire field.

Some of the shadowgrams obtained in this investigation are reproduced in Fig. 1. The one-eighth-inch steel spheres used were not permanently deformed by the impact. These shadowgrams

Fig. 1.

are of various shots taken at increasing depths of water penetration. The first shadowgram records the spherical missile just before impact; the last one records the entry after the shock wave has advanced eighteen sphere diameters from the surface. In the early stages of entry all the light passing through the shock wave region is strongly refracted and a complete shadow silhouette is recorded. At the later stages the light passes through with less deviation and some of the features of the interior of the region, such as cavity, can be observed. In the air above the water one finds the shadow of the splash which rises with supersonic speeds in air.

Information on the pressure field in the vicinity of the missile and shock wave can be augmented by obtaining shadowgrams using narrow $\frac{1}{8}$-inch spark beams. If the region between the shock front and the water surface is regarded as half a spherical lens, then this exploratory beam method shows that the "lens" has both a diverging and a converging action. The converging lens action occurs for rays which are nearly tangential to the shock front. These rays are refracted toward the center and produce the pattern of lines found in the shadow of the cavity. The diverging action occurs for rays striking the central portion of the shock wave region. In the early penetration stages these rays diverge so strongly that the shadow region is left completely dark and in the intermediate stages the downward refraction of these rays leaves a dark band below the water surface. A study of the refraction of individual spark beams shows that the density (or pressure) increases rapidly in going from the water surface toward the nose of the missile and also that the pressure increases abruptly in approaching the shock wave from the undisturbed region in front. These pressure gradients are quite large, since the distances are small; the peak pressure or stagnation pressure is at least 20,000 atmospheres at entry.

It is also interesting to note that the missile is set into vibration at impact and secondary waves are produced throughout the missile's water flight. In the last shadowgram one can observe these secondary waves after they have been reflected in the surface.

This investigation was partially supported by the ONR.

Author Index to Volume 21

References with (L) are to Letters to the Editor

Subject Index to Volume 21

References with (L) are to Letters to the Editor.